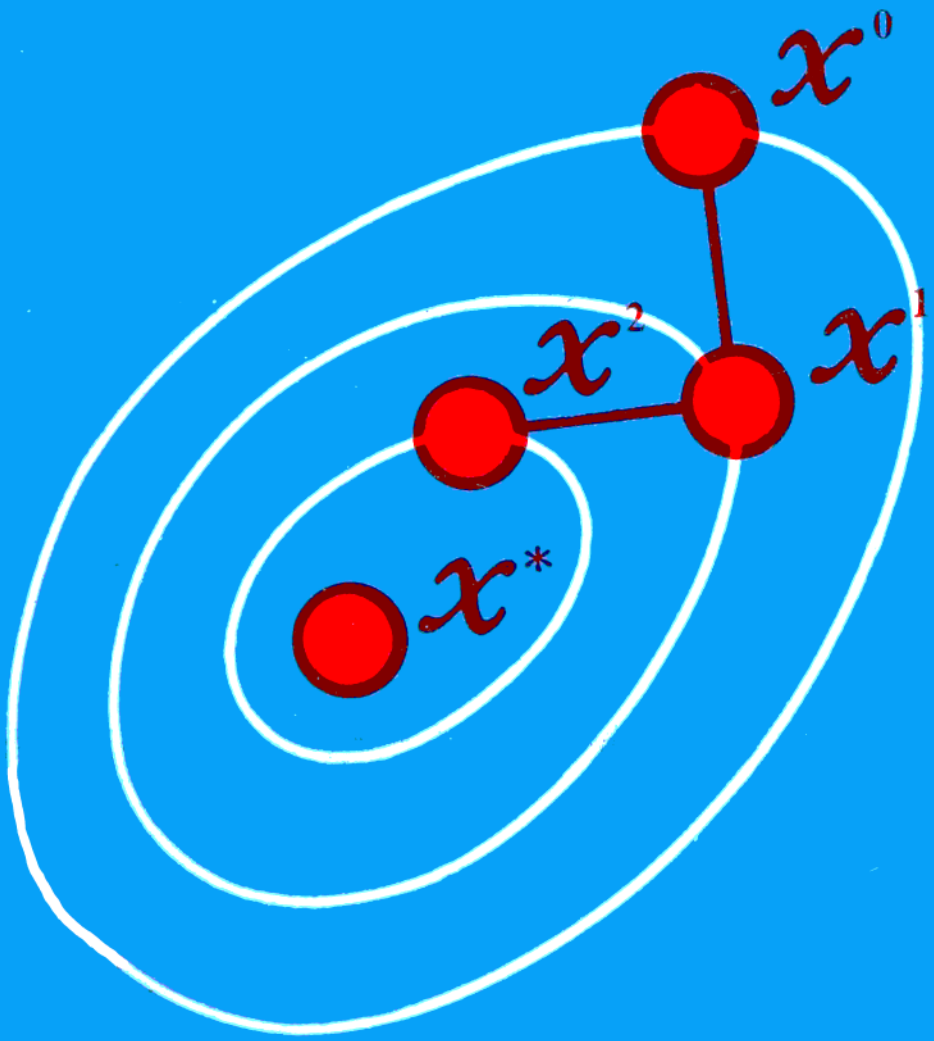
A. VOLKOV

NUMERICAL METHODS



MIR PUBLISHERS MOSCOW

Dr. E. Volkov (D. Sc. Physics and Mathematics) has presented the main numerical methods, that is the approximation of functions of one variable, integration, and the solution of nonlinear equations and systems of nonlinear equations. The solution of boundary-value problems for linear second-order ordinary differential equations is also considered. Finally Dr. Volkov covers the difference schemes for linear partial differential equations of the first and second order. This study aid closely follows *Higher Mathematics* (in four volumes) written by Ya. Bugrov and S. Nikolsky. Intended for engineering students.

NUMERICAL METHODS

Е. А. ВОЛКОВ

ЧИСЛЕННЫЕ МЕТОДЫ

«НАУКА» МОСКВА

E. A. VOLKOV

NUMERICAL METHODS

Translated from the Russian
by
Leonid Levant

MIR PUBLISHERS MOSCOW

First published 1986
Revised from the 1982 Russian edition

На английском языке

© Издательство «Наука». Главная редакция
физико-математической литературы, 1982

© English translation, Mir Publishers, 1986

CONTENTS

Preface		7
Introduction		9
Chapter 1.	**Polynomial Approximation to Functions**	18
Sec. 1.	Approximate Numbers and Operations with Them	19
Sec. 2.	Evaluating a Polynomial. Horner's Method	26
Sec. 3.	Taylor's Polynomials	28
Sec. 4.	Lagrange's Interpolation Polynomial	30
Sec. 5.	Linear Interpolation	35
Sec. 6.	Minimizing the Interpolation Error Estimate. Chebyshev Polynomials	36
Sec. 7.	Interpolation with Equally Spaced Points	42
Sec. 8.	Finite and Divided Differences	45
Sec. 9.	Newton's Interpolation Polynomial	49
Sec. 10.	Numerical Differentiation	53
Sec. 11.	Splines	62
Sec. 12.	Uniform Approximation of Functions	66
Sec. 13.	The Method of Least Squares	73
Sec. 14.	Analysis of Errors of Mean-square Approximations. "Smoothing" of Observations	88
Chapter 2.	**Numerical Integration**	99
Sec. 15.	Quadrature Formulas	99
Sec. 16.	Runge's Rule for a Practical Error Estimate	114
Sec. 17.	The Monte Carlo Method	118
Sec. 18.	Numerical Methods of Solving Cauchy's Problem for Ordinary Differential Equations	122
Chapter 3.	**Numerical Methods of Linear Algebra**	133
Sec. 19.	The Gauss Elimination Method	134
Sec. 20.	Norms and Condition of Matrices	145

Sec. 21.	The Method of Simple Iterations. Seidel's Method	150
Sec. 22.	The Sweep Method	155
Sec. 23.	Partial Eigenvalue Problems	160
Chapter 4.	**Methods of Solving Nonlinear Equations and Systems**	167
Sec. 24.	The Iteration Method	167
Sec. 25.	Newton's Method	179
Sec. 26.	The Method of Dividing a Line Segment into Two Equal Parts	184
Sec. 27.	The Method of Steepest (Gradient) Descent	185
Chapter 5.	**Methods of Solving the Boundary-value Problem for an Ordinary Linear Second-order Differential Equation**	187
Sec. 28.	Methods of Minimizing the Discrepancy. Galerkin's Method	187
Sec. 29.	The Difference Method. Basic Concepts of the Theory of Difference Schemes	194
Chapter 6.	**Difference Schemes for Partial Differential Equations**	209
Sec. 30.	Linear Equations with First-order Partial Derivatives	209
Sec. 31.	Mixed Problem for the Heat Equation	217
Sec. 32.	The Wave Equation	224
Sec. 33.	The Heat Equation in Two Spatial Variables	227
Sec. 34.	The Dirichlet Problem for Poisson's Equation	230
Subject Index		235

PREFACE

This book sets forth the foundations of numerical methods. It contains the material envisaged by the syllabus of the course "Higher Mathematics for Engineering Students". To understand almost the entire contents of this study aid, it is sufficient for the student to have a profound knowledge of the general course of higher mathematics contained in the following three textbooks by Professor Ya.S. Bugrov and Academician S.M. Nikolsky translated into English by Mir Publishers (Moscow) in 1982-83: (1) *Fundamentals of Linear Algebra and Analytical Geometry,* (2) *Differential and Integral Calculus,* and (3) *Differential Equations. Multiple Integrals. Series. Theory of Functions of a Complex Variable.*

A successful study of Secs. 14 and 17 requires a knowledge of random quantities from an elementary course of the theory of probability. Some additional notions used in the text are explained in the introduction and other sections whenever needed.

Chapter 1 is dedicated to the numerical methods of approximation of functions of one variable. Here, in addition to Taylor's polynomials, interpolation polynomials, and polynomials of best uniform approximation, a study of approximation by cubic splines is given. Considerable space is provided for the method of least squares in nonperiodic and periodic cases with analysis of the error of the method itself and the random error due to observation errors. The method is of great importance for engineering applications. Chapter 1 also includes a section on numerical differentiation used for construction of splines and Sec. 1 (introductory) in which approximate numbers are discussed.

Chapter 2 presents the numerical methods of integration. Along with the traditional quadrature formulas it gives a brief description of the Monte Carlo method for computing definite integrals. In Sec. 16 the reader will find a discussion of Runge's rule for a practical estimate of the error of quadrature formulas and also of the Richardson refinement of the result obtained. The numerical methods of integration of ordinary differential equations (the solution of Cauchy's problem) are discussed in Sec. 18.

Chapter 3 contains the numerical methods for solving the problems of linear algebra: the Gauss elimination method, the iteration method for solving systems of linear equations, and the methods for solving partial problems of the eigenvalues of matrices. Section 22 considers the sweep method for solving a system with a tridiagonal matrix (a three-point difference equation) which has been practised on a large scale.

Chapter 4 sets forth the basic approximate numerical methods for solving nonlinear equations and systems of nonlinear equations.

Chapter 5 is dedicated to the approximate methods for solving the boundary-value problem for an ordinary second-order linear differential equation. Section 28 gives the methods of minimizing the discrepancy and also the Galerkin method. Section 29 describes the difference method. On the basis of a two-point boundary-value problem, the fundamental concepts are introduced and the main points of the theory of difference schemes are explained.

Chapter 6 provides a study of difference schemes for linear differential equations with partial derivatives of first and second orders. Due consideration is given to the questions of approximation, stability, convergence, and economical efficiency of difference schemes.

I would like to express my appreciation to Academician S.M. Nikolsky and Professor Ya.S. Bugrov for their constant advice which has been of great help in writing this book.

I should also like to express my deep gratitude to the following reviewers for their critical evaluation of the manuscript at various stages: Corresponding Member of the USSR Academy of Sciences N.S. Bakhvalov, Professor V.A. Trenogin, and Assistant Professor N.A. Potapkov.

Evgenii A. Volkov

INTRODUCTION

In practice, in most cases we fail to find the exact solution of a mathematical problem. This happens mainly not because we do not know the way in which the exact solution is found but usually owing to the fact that the desired solution is not expressible in elementary or other functions usually known to us. Therefore numerical methods assume ever greater importance, especially in connection with the increasing role of mathematical methods in various fields of science and technology and owing to the appearance of high-efficient electronic computers.

By numerical methods we understand the methods of solving problems which are reduced to arithmetic and certain logical operations on numbers, that is, to the operations usually performed by a computer. Depending on the complicacy of a given problem, assigned accuracy, the method to be applied, and so forth, its solution may require a few dozen to many billions of operations. If the number of operations does not exceed a thousand, then one can successfully solve such a problem by using a desk piano-key computing machine without programmed control or a collection of tables of elementary functions. But one cannot manage without a high-speed computer if the solution of a problem requires, say, a million of operations to be performed and especially when the solution is to be found in the shortest possible time. For instance, the problems connected with twenty-four-hours' weather forecast must be solved in several hours, whereas in managing control of fast technological processes a solution is required to be found in fractions of a second.

A solution obtained by a numerical method is usually approximate, that is, it has some error. The following are the sources of the error of an approximate solution: (1) the lack of correspondence between a mathematical problem (mathematical model) and the real phenomenon under consideration; (2) the error of the initial data (input parameters); (3) the error of the method of solution; and (4) round-off errors in arithmetic and other-type operations on the numbers involved.

The error of a solution due to the first two sources is said to be *inherent*. This error may be present even if the solution of a posed

mathematical problem is found exactly. The question of the exactness with which a mathematical model describes this or that phenomenon is verified by comparing the results of experiments with the results of typical particular solutions for some values of the input parameters. We often succeeded in estimating the effect of the error of the initial data by elementary means, for instance, by varying the initial data within the bounds of their errors and by fixing the solution. If there are many initial data and their errors are of an accidental nature, then statistical methods may prove helpful. (Section 14 gives the estimates of a random error in approximating functions by the method of least squares.) In some cases, the inherent error may be regarded as the error of a function due to the error of its arguments. (The estimate of such an error based on Lagrange's formula is considered in Sec. 1.)

In most cases, numerical methods are approximate by themselves, that is, even if the initial data are void of errors and all the arithmetic operations are ideally performed, they yield the solution of the original problem with some error which is called the *error of the method* (applied). This is because a numerical method is usually applied to solve some other, simpler problem, which approximates the originally given problem. In a number of cases, the chosen numerical method is constructed on the basis of an infinite process which, in the limit, leads to the desired solution (for instance, computing the values of an elementary function with the aid of partial sums of a power series into which this function is expanded). But in reality, the passage to a limit usually fails to be realized, and the process interrupted at a certain step yields an approximate solution.

All sections of the present book dedicate considerable space to the analysis of the error of numerical methods. A numerical method usually depends on one or several parameters which may be controlled. Examples of such parameters include the number of iterations in solving systems of equations and the number of terms taken into account when summing a series, and also the step over which the values of the integrand are used in approximate evaluation of the definite integral. The error of the method applied or its estimate obtained usually depends on the appropriate parameter. One sometimes succeeds in obtaining the error estimate expressed only in terms of known quantities. With the aid of this estimate, it is possible to determine the values of the parameter specifying the method for which the error of the method lies within the required limits. But in most cases, the error estimate contains unknown constant multipliers, and the parameter of the method enters into it in the form of either a power or an exponential function. By such an estimate one can judge the rate of decrease of the error when changing the parameter of the method.

The rate of decrease of the error is an important characteristic of the method applied.

The most complicated thing is to take into account the round-off errors in arithmetic operations. If the number of operations to be performed is not large, then *round-off errors* in manual calculations can be taken into account by the method set forth in Sec. 1 dedicated to the elementary theory of errors for a small number of operations. Two situations may occur when solving problems on a computer. If the number of arithmetic operations to be performed is small, then round-off errors are usually not revealed, since numbers are put into a computer with ten or even more significant decimal digits, whereas it is seldom required that the final result include more than five significant digits. If a problem is complicated (for instance, a problem reduced to the solution of partial differential equations), and its approximate solution requires, say, 10^7 arithmetic operations, then it is unrealistic in this case to take into account the effect of round-off errors in each operation. For such an account we shall obtain a very excessive estimate of the error corresponding to the most unfavourable case. And round-off errors behave in a rather random way as far as their magnitude and sign are concerned. Therefore there are prerequisites for their mutual compensation. But all the same, one has to take into consideration the errors due to rounding off when solving complicated problems.

One and the same problem can be solved by different approximate methods. The sensitivity to round-off errors depends essentially on the numerical method chosen. Chapters 5 and 6 analyze the difference methods of solving the boundary-value problems for differential equations and pay attention to the so-called stable methods (difference schemes) which are of low sensitivity to round-off errors. Among them are convergent iteration methods since the appearing errors are corrected at the subsequent iterations.

A numerical method may be regarded as successfully chosen if its error is a fraction of the inherent error, and the error due to rounding off, called the *computational error,* is at least several times less than the error of the method. If there is no inherent error, then the error of the method must be somewhat less than the assigned accuracy of solution.

Besides the requirement to achieve the preassigned accuracy, the chosen numerical method should meet a number of other demands. Preference is given to the method which is realized through a minimum number of operations, which requires a less powerful storage unit fitted with the computer, and, finally, which is logically simpler and, thus, is realized more rapidly on a computer. The enumerated conditions usually

contradict one another, therefore, when deciding which numerical method is to be used, a compromise should be made.

Let us introduce some notions of mathematical and functional analysis which are widely used in modern literature on numerical methods. The reader may grasp these notions as they encounter in the main text.

1. Let $\varphi(h)$ be some function of the variable h with a finite domain of definition D_φ on the semi-axis $h > 0$, and $h \in D_\varphi$ can attain arbitrarily small values. Then, if there are positive numbers h_0, c, and k such that for all $h \in D_\varphi$ satisfying the condition $0 < h \leqslant h_0$ the inequality

$$|\varphi(h)| \leqslant ch^k$$

is fulfilled, we write

$$\varphi(h) = O(h^k)$$

and say that $\varphi(h)$ *is a capital O of* h^k *(as* $h \to 0$*).*

According to the given definition, the following obvious properties are fulfilled. If $\varphi(h) = O(h^k)$, $\psi(h) = O(h^k)$, and $D_\varphi = D_\psi$, then $\varphi(h) + \psi(h) = O(h^k)$, i.e. $O(h^k) + O(h^k) = O(h^k)$. If $k > m > 0$, then $O(h^k)$ is at the same time $O(h^m)$. Finally, if $\varphi(h) = O(h^k)$, then $a\varphi(h) = O(h^k)$, where a is a constant independent of h.

Example. $\sin^2 2h = O(h^2)$, since $\sin^2 2h \leqslant 4h^2 \; \forall \, h$.

Let there be given now a function $\varphi(h, \tau)$ of the positive arguments h and τ which can attain arbitrarily small values. Then if there are positive numbers h_0, τ_0, c, k, and m such that for all permissible values of h and τ satisfying the conditions $0 < h \leqslant h_0$ and $0 < \tau \leqslant \tau_0$ the inequality

$$|\varphi(h, \tau)| \leqslant c(h^k + \tau^m)$$

is fulfilled, we write

$$\varphi(h, \tau) = O(h^k + \tau^m)$$

and say that $\varphi(h, \tau)$ *is a capital O of* $h^k + \tau^m$ *(as* $h \to 0, \tau \to 0$*).*

Let us introduce one more analogous notion.

The function $\Phi(N)$ defined for all natural $N > N_0 > 0$ *is a capital O of* N^k *(as* $N \to \infty$*),* that is, $\Phi(N) = O(N^k)$ if there is a constant $c > 0$ such that for all natural $N > N_0$

$$|\Phi(N)| \leqslant cN^k.$$

2. We shall say that a function $f(x)$ *belongs to the class* $C_k[a, b]$ and write $f \in C_k[a, b]$ if the function f is defined on the interval $[a, b]$ and has continuous derivatives up to order k inclusive on it. This means that k

times continuously differentiable function f^* exists on some interval (A, B) containing the interval $[a, b]$. This function coincides with f on $[a, b]$. By the values of the indicated derivatives of the function f at the end points of the interval $[a, b]$ we understand the values of the corresponding derivatives of the function f^*.

Example. Let there be given the function $f(x) = x^{5/2}$ on the interval $[0, 1]$. Consider the function $f^*(x) = |x|^{5/2}$ on the interval $(-1, 2)$. Obviously, the interval $[0, 1]$ is contained in the interval $(-1, 2)$, the function f^* is twice continuously differentiable on the interval $(-1, 2)$, namely,

$$\frac{df^*(x)}{dx} = \frac{5}{2}|x|^{3/2}\operatorname{sgn} x,$$

$$\frac{d^2f^*(x)}{dx^2} = \frac{15}{4}|x|^{1/2}.$$

In addition, $f^*(x) = f(x)\ \forall\, x \in [0, 1]$. Consequently, $f \in C_2[0, 1]$, and $f'(x) = (5/2)x^{3/2}$, $f''(x) = (15/4)x^{1/2}$, $x \in [0, 1]$.

For $k = 0$, instead of $C_0[a, b]$, we shall use the symbol $C[a, b]$. The notation $f \in C[a, b]$ means that the *function f is continuous on the interval* $[a, b]$.

3. Given in differential calculus are the notions of partial derivatives of a function of several variables at the interior points of the domain of its definition.

To formulate appropriate problems, we should extend these notions to a closed domain including the boundary points.

Let the function f be given on the closed domain $\bar{G}$ of the n-dimensional space $\mathbf{R}^n$, $n \geqslant 2$. We shall say that *f has a continuous partial derivative on the closed domain $\bar{G}$* if a function f^* which coincides with f on $\bar{G}$ and has a corresponding continuous partial derivative on G^* exists on some open domain G^* which contains the closed domain $\bar{G}$. In this case, we shall take the values of the corresponding partial derivative of the function f^* for the values of the considered partial derivative of the function f at the boundary points of the domain G.

Definition. We shall say that the function *f is continuously differentiable k times on the closed domain $\bar{G}$,* that is, *belongs to the class $C_k(\bar{G})$*, and write $f \in C_k(\bar{G})$ if the function f has all the continuous partial derivatives up to order k inclusive on $\bar{G}$.

Example. Let there be given the function $f(x_1, x_2) = x_2 \cos (x_1 - 2)^{1/2}$ on the closed square $\bar{G} = \{(x_1, x_2)\colon 0 \leqslant x_i \leqslant 1, i = 1, 2\}$. Let us define on the open circle $G^* = \{(x_1, x_2)\colon x_1^2 + x_2^2 < 4\}$ the function $f^*(x_1, x_2)$ by the same formula: $f^*(x_1,$

$x_2) = x_2 \cos (x_1 - 2)^{1/2}$. Obviously, $\bar{G}$ is contained in G^*, $f^*(x_1, x_2)$ coincides with $f(x_1, x_2)$ on $\bar{G}$, and the function $f^*(x_1, x_2)$ is continuously differentiable k times on G^* for any k. According to the above definition, $f \in C_k(\bar{G})$ $\forall$ k.

4. Let M be the set of elements f, g, r, ... of an arbitrary nature.

A set M is called a *metric space* if any pair of its elements is associated with a nonnegative number $\rho(f, g)$, called the *distance* between the elements f and g, which satisfies the *following axioms of distance:*

(1) $\rho(f, g) = 0$ *if and only if* $f = g$;

(2) $\rho(f, g) = \rho(g, f)$ $\forall f, g \in M$;

(3) $\rho(f, r) \leqslant \rho(f, g) + \rho(g, r)$ $\forall f, g, r \in M$.

Axiom (3) is called the *triangle inequality*. The function $\rho(f, g)$ will also be called the *metric* of the space M.

A metric space may be exemplified by any set M of the points $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, ... of the n-dimensional space $\mathbf{R}^n$ with the distance

$$\rho(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. The fulfilment of Axioms (1) and (2) for the given distance is obvious. The triangle inequality follows from *Minkowski's inequality* (proved by linear algebra):

$$|\mathbf{x} + \mathbf{y}| \leqslant |\mathbf{x}| + |\mathbf{y}|,$$

where $\mathbf{x}$ and $\mathbf{y}$ are arbitrary points of the space $\mathbf{R}^n$, $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n)$, and $|\mathbf{x}| = (x_1^2 + \ldots + x_n^2)^{1/2}$. Indeed,

$$\rho(\mathbf{x}, \mathbf{z}) = |\mathbf{x} - \mathbf{z}| = |(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{z})|$$
$$\leqslant |\mathbf{x} - \mathbf{y}| + |\mathbf{y} - \mathbf{z}| = \rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z}).$$

In particular, the set of points forming the closed unit ball $\bar{S} = \{\mathbf{x}: |\mathbf{x}| \leqslant 1\}$ is a metric space with the distance indicated above.

5. A set M is called a *linear space* if the operations of addition and multiplication by the real numbers within the bounds of M and satisfying the following conditions are defined in this set:

(1) *addition is associative:* $(f + g) + r = f + (g + r)$;
(2) *commutative:* $f + g = g + f$;
(3) *there exists a zero element* $\theta \in M$, *i.e.* $f + \theta = f$ $\forall f \in M$;
(4) $0 \cdot f = \theta$ $\forall f \in M$;
(5) $(\alpha + \beta)f = \alpha f + \beta f$;

(6) $\alpha(f + g) = \alpha f + \alpha g$;
(7) $\alpha(\beta f) = (\alpha\beta)f$;
(8) $1 \cdot f = f$;

here α and β are *real numbers.*

6. A system of elements $f_1, f_2, \ldots, f_n$ of the linear space M is said to be *linearly dependent* if there are numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$, not all zero, such that

$$\alpha_1 f_1 + \alpha_2 f_2 + \ldots + \alpha_n f_n = \theta.$$

If this equality is possible only for $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$, then the system of elements $f_1, f_2, \ldots, f_n$ is said to be *linearly independent.*

7. A set F is called a *linear normed space* if it is linear and each element $f \in F$ is associated with a real number $\|f\|$ which is called the *norm* of f and satisfies the *axioms of norm:*

(1) $\|f\| \geqslant 0$, *and* $\|f\| = 0$ *if and only if* $f = \theta$, *that is,* f *is a zero element in* F;
(2) $\|\alpha f\| = |\alpha| \, \|f\|$ *for any real* α;
(3) $\|f + g\| \leqslant \|f\| + \|g\| \; \forall f, g \in F$.

Axiom (3) is called the *triangle inequality* for the norm.

Consider an example of a normed space. The class $C[a, b]$ of all continuous functions specified on the interval $[a, b]$ is, obviously, a linear space, since the sum of any two continuous functions is continuous, and a continuous function multiplied by any number is also continuous. The zero element in this space is represented by the only function which is identically equal to zero on $[a, b]$.

If we introduce the norm

$$\|f\| = \max_{[a, b]} |f(x)| \tag{1}$$

into the class $C[a, b]$, then it becomes a normed space. The fulfilment of Axioms (1) and (2) for the norm introduced is obvious. Let us verify the triangle inequality. We have

$$\|f + g\| = \max_{[a, b]} |f(x) + g(x)|.$$

Since $|f(x) + g(x)|$ is a continuous function on the interval $[a, b]$, according to Weierstrass' theorem, there is a point $x^* \in [a, b]$ such that

$$\max_{[a, b]} |f(x) + g(x)| = |f(x^*) + g(x^*)|.$$

Hence, there readily follows the triangle inequality:

$$\|f + g\| = |f(x^*) + g(x^*)| \leqslant |f(x^*)| + |g(x^*)|$$
$$\leqslant \max_{[a, b]} |f(x)| + \max_{[a, b]} |g(x)| = \|f\| + \|g\|.$$

8. *Any linear normed space is at the same time a metric space with the distance (metric)*

$$\rho(f, g) = \|f - g\|. \tag{2}$$

Indeed, Axioms (1) and (2) for the distance introduced are fulfilled by virtue of Axioms (1) and (2) for the given norm for $\alpha = -1$*. The triangle inequality for the distance (2) follows from the triangle inequality for the norm:

$$\rho(f, r) = \|f - r\| = \|(f - g) + (g - r)\|$$
$$\leqslant \|f - g\| + \|g - r\| = \rho(f, g) + \rho(g, r).$$

9. In one and the same linear space, the norm can be introduced by different methods. For instance, in the class of continuous functions $C[0, 1]$ the norm can also be specified, besides the method (1), in the following form:

$$\|f\| = \left(\int_0^1 f^2(x)dx \right)^{1/2}. \tag{3}$$

We are not going to dwell on the verification of the axioms for this norm, postponing this question till Sec. 13.

In order to distinguish between various norms, we shall use appropriate subscripts. For example, for the norm (1) the following symbols are used: $\|f\|_{C[a, b]}$ and $\|f\|_C$, and for the norm (3) the symbol $\|f\|_{L_2}$ is frequently utilized.

Depending on the norm introduced in one and the same linear space, various properties of the obtained normed space may undergo a change and, in particular, the "physical" meaning of the metric generated by the norm is changed. For instance, the distance (2) specified with the aid of the norm (1), that is,

$$\rho(f, g) = \|f - g\|_C = \max_{[a, b]} |f(x) - g(x)|, \tag{4}$$

---

* The axioms of distance are formulated in Item 4, and the axioms of norm in Item 7.

means the maximum-modulus deviation of the functions f and g from each other on the interval $[a, b]$ and, when using the norm (3), the distance has the form

$$\rho(f, g) = \|f - g\|_{L_2} = \left(\int_0^1 (f(x) - g(x))^2 dx\right)^{1/2} \tag{5}$$

and bears the sense of the mean-square deviation of the functions f and g.

The choice of the norm is often dictated by the conditions of a concrete problem. For instance, when considering uniform approximations to functions (Sec. 12), the norm (1) is needed, and when approximating functions by the method of least squares (Sec. 13), the use of the norm of the type (3) naturally suggests itself.

10. In metric and normed spaces, the notion of the convergence of the sequence of their elements is introduced in a similar way as it is done in the n-dimensional space $\mathbf{R}^n$.

Definition. The sequence $\{f_n\}$ of elements of a metric (normed) space is said to converge to its element f if $\rho(f, f_n) \to 0$ ($\|f - f_n\| \to 0$) as $n \to \infty$.

The convergence in a metric space is called the *convergence in metric,* and in a normed space—the *convergence in norm.* The convergence in the metric (4) or, which is the same, in the norm (1) is *uniform,* while the convergence in the metric (5) or in the norm (3) is known as the *convergence in the mean-square sense* or, simply, as the *convergence in the mean.*

Approximate solutions of many problems, as well as their exact solutions, may be conveniently treated as elements of metric or normed spaces. Then the error is measured by the distance or the norm of the difference between the exact and approximate solutions in the appropriate space.

Chapter 1

POLYNOMIAL APPROXIMATION TO FUNCTIONS

Most numerical methods are based on replacing complicated objects, equations, etc. by simpler ones. One of the fundamental notions in mathematics is that of the function. An algebraic polynomial is the most convenient function to be handled in practice. To define a polynomial, it is only necessary to specify a finite number of its coefficients. Algebraic polynomials are readily evaluated, differentiated, integrated, and so forth. Therefore they are widely used for approximating various functions.

The present chapter deals with the practical methods of approximating functions of one variable on an interval. In particular, considered here are approximations effected with the aid of the algebraic polynomials of four types: (1) Taylor's polynomials, (2) interpolation polynomials, (3) polynomials of best uniform approximation and polynomials close to them, and (4) polynomials of best mean-square approximation (found by the method of least squares).

This chapter also includes Sec. 10 which is dedicated to numerical differentiation. It contains simplest formulas of approximate numerical differentiation using the values of function at a finite number of points, and dwells on approximating the corresponding derivatives of a function on an interval with the aid of the derivatives of an interpolation polynomial.

Section 11 describes the approximations by cubic splines (that is, by smooth functions composed of a finite number of algebraic polynomials of third degree).

Along with algebraic polynomials, the method of least squares (Sec. 13) also uses trigonometric polynomials which are more natural for approximating periodic functions. Section 14 considers the error of this method and describes the effect of random errors in discrete values of a function (observation errors) on the approximating polynomial found by the least-squares method. In particular, the smoothing of observations is discussed.

Sec. 1. Approximate Numbers and Operations with Them

Let a be the exact, generally speaking, unknown numerical value of a quantity, and let a^* be a known approximate numerical value of this quantity (an *approximate number*). The number

$$\Delta(a^*) = |a - a^*| \tag{1}$$

is called the *absolute error* of the approximate number a^*, and the quantity

$$\delta(a^*) = \frac{\Delta(a^*)}{|a^*|} \tag{2}$$

is termed its *relative error.*

Any number $\bar{\Delta}(a^*)$ ($\bar{\delta}(a^*)$) of which it is known that

$$\Delta(a^*) \leqslant \bar{\Delta}(a^*) \quad (\delta(a^*) \leqslant \bar{\delta}(a^*)) \tag{3}$$

is called the *limiting absolute (limiting relative) error* of the approximate number a^*. For the sake of definiteness, we shall regard that if $a^* \neq 0$ and the numbers $\bar{\Delta}(a^*)$ and $\bar{\delta}(a^*)$ are known, then they are connected by the relationship $\bar{\delta}(a^*) = \bar{\Delta}(a^*)/|a^*|$.

Taking Account of Errors in Arithmetic Operations. It is obvious that if $c = a + b$, $c^* = a^* + b^*$ or $c = a - b$, $c^* = a^* - b^*$, then

$$\Delta(c^*) \leqslant \Delta(a^*) + \Delta(b^*) \tag{4}$$

and for $\bar{\Delta}(c^*)$ it is natural to take

$$\bar{\Delta}(c^*) = \bar{\Delta}(a^*) + \bar{\Delta}(b^*). \tag{5}$$

Thus, *as two approximate numbers are added or subtracted, their limiting absolute errors are added together.* This rule is valid for the algebraic addition of any finite number of approximate numbers.

Let $u = ab$, $u^* = a^*b^*$, $v = a/b$, and $v^* = a^*/b^*$. We have

$$\begin{aligned}\Delta(u^*) = |u - u^*| &= |ab - a^*b^*| \\ &= |ab - a^*b + a^*b - a^*b^*| \\ &\leqslant |b|\,|a - a^*| + |a^*|\,|b - b^*| \\ &\leqslant (|b^*| + \Delta(b^*))\Delta(a^*) + |a^*|\Delta(b^*),\end{aligned}$$

that is, the absolute error of the product of two approximate numbers a^* and b^* satisfies the inequality

$$\Delta(u^*) \leqslant |b^*|\Delta(a^*) + |a^*|\Delta(b^*) + \Delta(a^*)\Delta(b^*). \tag{6}$$

Analogously, under the condition that $|b^*| > \Delta(b^*)$ and consequently, $b \neq 0$, we obtain

$$\Delta(v^*) = \left|\frac{a}{b} - \frac{a^*}{b^*}\right| = \left|\frac{ab^* - a^*b}{bb^*}\right|$$

$$= \left|\frac{ab^* - a^*b^* + a^*b^* - a^*b}{bb^*}\right| \leqslant \frac{|b^*|\Delta(a^*) + |a^*|\Delta(b^*)}{(|b^*| - \Delta(b^*))|b^*|}.$$

Hence, since $(b^*) - \Delta(b^*) = (1 - \delta(b^*))|b^*| > 0$, we arrive at the following inequality for the absolute error of the quotient of two approximate numbers:

$$\Delta(v^*) \leqslant \frac{|b^*|\Delta(a^*) + |a^*|\Delta(b^*)}{(1 - \delta(b^*))|b^*|^2}. \tag{7}$$

Let $a^* \neq 0$ and $|b^*| > \Delta(b^*)$. Dividing the inequality (6) by the product $|a^*||b^*|$, and the inequality (7) by the ratio $|a^*|/|b^*|$, we obtain the inequalities which are satisfied by the relative errors of the product and quotient:

$$\delta(u^*) \leqslant \delta(a^*) + \delta(b^*) + \delta(a^*)\delta(b^*), \tag{8}$$

$$\delta(v^*) \leqslant \frac{\delta(a^*) + \delta(b^*)}{1 - \delta(b^*)}. \tag{9}$$

For each of the inequalities (4) and (6)-(9), it is possible to give examples in which $\Delta(a^*) \neq 0, \Delta(b^*) \neq 0$, and the corresponding inequality turns into an equality. On the other hand, it may happen that $\Delta(a^*) \neq 0$, $\Delta(b^*) \neq 0$, and the left-hand member of some of the inequalities (4) and (6)-(9) is actually equal to zero.

By virtue of the inequalities (6)-(9), we may set the following quantities as the limiting absolute and limiting relative errors of a product and a quotient:

$$\bar{\Delta}(u^*) = |b^*|\bar{\Delta}(a^*) + |a^*|\bar{\Delta}(b^*) + \bar{\Delta}(a^*)\bar{\Delta}(b^*), \tag{10}$$

$$\bar{\Delta}(v^*) = \frac{|b^*|\bar{\Delta}(a^*) + |a^*|\bar{\Delta}(b^*)}{(1 - \bar{\delta}(b^*))|b^*|^2}, \tag{11}$$

$$\bar{\delta}(u^*) = \bar{\delta}(a^*) + \bar{\delta}(b^*) + \bar{\delta}(a^*)\bar{\delta}(b^*), \tag{12}$$

$$\bar{\delta}(v^*) = \frac{\bar{\delta}(a^*) + \bar{\delta}(b^*)}{1 - \bar{\delta}(b^*)}. \tag{13}$$

In the formulas (11) and (13) it is assumed that $\bar{\delta}(b^*) < 1$.

Since the limiting relative errors $\bar{\delta}(a^*)$, $\bar{\delta}(b^*)$ and, consequently, the relative errors $\delta(a^*)$, $\delta(b^*)$ are usually considerably smaller than unity, instead of the formulas (10)-(13), use is made of the approximate formulas

$$\bar{\Delta}(u^*) \approx |b^*|\bar{\Delta}(a^*) + |a^*|\bar{\Delta}(b^*), \tag{14}$$

$$\bar{\Delta}(v^*) \approx \frac{|b^*|\bar{\Delta}(a^*) + |a^*|\bar{\Delta}(b^*)}{|b^*|^2}, \tag{15}$$

$$\bar{\delta}(u^*) \approx \bar{\delta}(a^*) + \bar{\delta}(b^*), \tag{16}$$

$$\bar{\delta}(v^*) \approx \bar{\delta}(a^*) + \bar{\delta}(b^*). \tag{17}$$

Thus, *as approximate numbers are added or subtracted,* according to (5), *their limiting absolute errors are added together, and as approximate numbers are multiplied or divided,* according to (16) and (17), *their limiting relative errors are added together.*

The Error of a Function. For the sake of definiteness, consider the function of two variables $f(x, y)$ which is continuously differentiable in a certain domain G. Suppose that the numbers x_0^* and y_0^* are approximate values of the coordinates of some point (x_0, y_0), and the closed rectangle

$$\bar{R} = \{(x, y): |x - x_0^*| \leqslant \bar{\Delta}(x_0^*), |y - y_0^*| \leqslant \bar{\Delta}(y_0^*)\},$$

containing both points (x_0^*, y_0^*) and (x_0, y_0), is situated entirely in the domain G.

Let us designate $z_0 = f(x_0, y_0)$ and $z_0^* = f(x_0^*, y_0^*)$. According to Lagrange's formula of finite increments, we have

$$z_0 - z_0^* = f_x'(\xi, \eta)(x_0 - x_0^*) + f_y'(\xi, \eta)(y_0 - y_0^*), \tag{18}$$

where ξ and η are the coordinates of some point of the line segment connecting the points (x_0, y_0) and (x_0^*, y_0^*). Hence

$$\Delta(z_0^*) = |z_0 - z_0^*| \leqslant |f_x'(\xi, \eta)|\Delta(x_0^*) + |f_y'(\xi, \eta)|\Delta(y_0^*). \tag{19}$$

Let c_1 and c_2 be certain estimates for $|f_x'|$ and $|f_y'|$ on the rectangle $\bar{R}$, that is,

$$\max_{\bar{R}}|f_x'| \leqslant c_1, \quad \max_{\bar{R}}|f_y'| \leqslant c_2. \tag{20}$$

It is desirable to have the estimates c_1 and c_2 as small as possible.

Since $(\xi, \eta) \in \bar{R}$, taking into account (3), (19), and (20), we may take the quantity

$$\bar{\Delta}(z_0^*) = c_1\bar{\Delta}(x_0^*) + c_2\bar{\Delta}(y_0^*) \tag{21}$$

as the *limiting absolute error of the value of the function* $z_0^* = f(x_0^*, y_0^*)$.

Sometimes, we proceed not in a rigorous way by supposing that

$$\bar{\Delta}(z_0^*) \approx |f_x'(x_0^*, y_0^*)|\bar{\Delta}(x_0^*) + |f_y'(x_0^*, y_0^*)|\bar{\Delta}(y_0^*). \tag{22}$$

The inverse problem is frequently encountered, namely, we are given the condition that the absolute error $\Delta(z_0^*)$ of the value of a function does not exceed some number $\varepsilon > 0$, that is,

$$\Delta(z_0^*) \leqslant \varepsilon. \tag{23}$$

It is required to find the constraints to be imposed on the errors of the arguments.

Let us now assume in addition that the domain G in which a differentiable function $f(x, y)$ is given is convex, that is, any two points of the domain G can be joined by a straight-line segment contained entirely in G. For instance, the domain G is an open circle or rectangle. Let also

$$\sup_G |f_x'| \leqslant C_1, \quad \sup_G |f_y'| \leqslant C_2 \tag{24}$$

and, as before, x_0^* and y_0^* are approximate values of the coordinates of some point (x_0, y_0), and it is supposed that both points (x_0, y_0) and (x_0^*, y_0^*) are contained in the domain G.

Since the domain G is convex, the difference between the values of the function f at the indicated points can be represented by Lagrange's formula (18) from which the inequality (19) follows. Consequently, since $(\xi, \eta) \in G$, for the condition (23) to be fulfilled, it is sufficient that the absolute errors $\Delta(x_0^*)$ and $\Delta(y_0^*)$ of the arguments satisfy the inequality

$$C_1\Delta(x_0^*) + C_2\Delta(y_0^*) \leqslant \varepsilon. \tag{25}$$

The direct and inverse problems of obtaining the estimates of the errors of differentiable functions of one or several variables are solved in a similar way.

The Rules for Writing Approximate Numbers. Let the approximate number a^* be given in the following notation:

$$a^* = \pm\alpha_n\alpha_{n-1}\cdots\alpha_0\cdot\alpha_{-1}\cdots\alpha_{-m},$$

where α_j are decimal digits $\left(a^* = \pm \sum_{j=-m}^{n} \alpha_j 10^j\right)$.

The first nonzero digit of a given number on the left and all the digits found on the right are said to be *significant*. For instance, the numbers 25.047 and -0.00250 have 5 and 3 significant digits, respectively.

A digit α_j is said to be *correct* if $\Delta(a^*) \leqslant 10^j$, that is, if the absolute error of the number a^* does not exceed one unity of the appropriate digit place (position) of a decimal number.

If an approximate number is to be written without indicating its absolute, or, speaking more exactly, its limiting absolute error, then only its correct digits are written out. In this event, the correct zeros on the right of the number are not rejected. For instance, the numbers 0.0344 and 0.034400 are different approximate numbers. Speaking of the first number, we may only assert that its absolute error does not exceed 0.0001, whereas the notation of the second number implies that its absolute error is not greater than 10^{-6}.

If the integral part of an approximate number has more significant digits than correct digits, then we resort to the notation in normalized form. For example, $a^* = 0.390 \cdot 10^5$. From this notation it is clear that the number a^* has three correct significant digits. In this case, the notation of the form $a^* = 39{,}000$ is inadmissible.

The aforegoing approximate numbers can also be rendered in normalized form: $0.344 \cdot 10^{-1}$ and $0.34400 \cdot 10^{-1}$.

Frequent use is made of the notation of the form

$$a = a^* \pm \bar{\Delta}(a^*),$$

which means that the unknown quantity a satisfies the inequalities

$$a^* - \bar{\Delta}(a^*) \leqslant a \leqslant a^* + \bar{\Delta}(a^*).$$

Here, the *quantity $\bar{\Delta}(a^*)$ is written out with one or two significant digits, and the less significant digit (position) in a^* corresponds to that in $\bar{\Delta}(a^*)$.* For example,

$$a = 2.730 \pm 0.017. \tag{26}$$

The notations

$$a = 2.73018 \pm 0.017,$$

$$a = 2.73018 \pm 0.01592 \tag{27}$$

are unnatural.

If, by hypothesis, it is known that some number is exact, say, the computation step $h = 0, 1$, then the correct zeros on the right are not written out.

We may speak of the number of correct significant digits in an approximate number and of the number of correct digits after the decimal point. As a rule, in practical computations approximate numbers have some digits after the decimal point, that is, they have a fractional part. For example, the approximate number $a^* = 25.030$ has five correct significant digits and three correct digits after the decimal point, while the number $b^* = 0.00404$ has, vice versa, three correct significant digits and five correct digits after the decimal point.

Obviously, *the absolute error of an approximate number is completely characterized by the number of correct digits after the decimal point, whereas its relative error is characterized by the number of correct significant digits.*

Rounding Off Numbers. In carrying out various computations, there often arises the necessity of rounding off numbers, that is, of decreasing the number of decimal places.

The Rule for Rounding Off Numbers. *When the first digit dropped is less than 5, the preceding digit is not changed; when the first digit dropped is greater than or equal to 5, and some succeeding digit is not zero, the preceding digit is increased by 1; when the first digit dropped is 5, and all the succeeding digits are zero, the commonly accepted rule (computer's rule) is to make the preceding digit even, that is, add 1 to it if it is odd, and leave it alone if it is already even.* For instance, 2.324, 2.316, and 2.315 would take the form 2.32 if rounded off to two places.

Example. Round off, respectively, to two, three, and four decimal places the following numbers: 3.14159, -0.0025, 84.009974.

Answer. 3.14, -0.003, 84.0100.

Obviously, *the round-off error does not exceed (in absolute value) half the unity of the less significant digit (position) left.*

A repeated rounding off is not recommended, since it may lead to an increased error. For instance, rounding off the number 18.34461 first to three and then to two decimal places, we obtain 18.345 and 18.35, respectively. When rounding off at once to two decimal places, we have 18.34. The absolute error of the repeated rounding off turns out to be equal to 0.00539, while in one-time rounding off the absolute error is equal to 0.00461.

When rounding off an approximate number, its limiting absolute error is increased owing to the round-off error. For instance, when passing from the notation (27) to the notation (26), where the number 2.73018 is rounded off to three places after the decimal point, its limiting absolute error 0.01592 is increased to reach the magnitude $0.017 > 0.01592 + 0.00018 > 0.016$ also represented by three digits after the decimal point.

It is common practice to round off constants known with a large number of digits, products of multidigit numbers, quotients, etc. For instance, when multiplying two approximate numbers each having six correct significant digits, we obtain the product with eleven or twelve digits. The relative error of the product may turn out to be twice the error of the cofactors, since the inequality (8) may turn into an equality. Therefore the product has approximately six correct significant digits, the remaining

digits, as a rule, yielding no useful information. In this case, it is only natural to round off the obtained product to six significant digits.

There are the following rules for arithmetic operations with approximate numbers:

When multiplying or dividing approximate numbers having, generally speaking, different numbers of correct significant digits, the result obtained is rounded off to a number of digits coinciding with the minimum number of correct significant digits in the original numbers.

When adding or subtracting approximate numbers having an equal number of correct digits after the decimal point, the result obtained is not to be rounded off. When adding or subtracting approximate numbers with a different number of correct digits after the decimal point, the result obtained is to be rounded off to the number of digits coinciding with the minimum number of correct digits after the decimal point in the original numbers.

Remark 1. When carrying out manual computations, in order to reduce the effect of round-off errors in approximate numbers, it is usual practice, in addition to correct significant digits, to leave one or two spare digits and to proceed according to the above-formulated rules, taking into account the spare digits. The latter are rejected when rounding off the final result.

Remark 2. When subtracting numbers close to each other in magnitude, considerable number of significant digits may be lost, which would lead to a less accurate result. For example, let it be required to evaluate $\sqrt{543} - \sqrt{540}$, where 543 and 540 are exact numbers. We have $\sqrt{543} = 23.30...$ and $\sqrt{540} = 23.23...$. Rounding off these numbers to three significant digits, we arrive at the result containing only one significant digit: $\sqrt{543} - \sqrt{540} \approx 23.3 - 23.2 = 0.1$.

Let us now get rid of subtracting close approximate numbers:

$$\sqrt{543} - \sqrt{540} = \frac{3}{\sqrt{543} + \sqrt{540}} \approx \frac{3}{23.3 + 23.2} = \frac{3}{46.5} = 0.06451....$$

Rounding off the obtained number to three significant digits,we get $\sqrt{543} - \sqrt{540} \approx 0.0645$.

With the aid of more accurate computations, we can show that all the three significant digits in the last result are correct, although, as in the preceding case, the approximate values of the radicals $\sqrt{543}$ and $\sqrt{540}$ were used with three significant digits.

This example shows that, whenever it is possible, *one should avoid the subtraction of close approximate numbers.* And if it is impossible, one should carry out the intermediate computations with a higher accuracy, taking into account the loss of significant digits in subtraction.

Sec. 2. Evaluating a Polynomial. Horner's Method

Consider the algebraic polynomial

$$P_n(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1} + a_nx^n, \tag{1}$$

where $a_0, \ldots, a_n$ are numerical coefficients, and n is the degree of the polynomial. To evaluate the polynomial (1) with a fixed $x = a$, we can proceed in different ways. For instance, first, with the aid of $n - 1$ multiplications, we find the powers of a, that is, $a, a^2, \ldots, a^n$. Then, in accordance with the formula (1), where $x = a$, performing n more multiplications and n additions, we get $P_n(a)$. This most natural, at the first glance, method requires, in the general case, the performance of $3n - 1$ arithmetic operations for $n \geqslant 1$.

However, the values of the polynomial are found more economically if it is rewritten in the following manner:

$$P_n(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-2} + x(a_{n-1} + xa_n))\ldots)). \tag{2}$$

According to the formula (2), the evaluation of $P_n(a)$ is reduced to the successive finding of the following quantities:

$$\begin{aligned} b_n &= a_n, \\ b_{n-1} &= a_{n-1} + ab_n, \\ b_{n-2} &= a_{n-2} + ab_{n-1}, \\ &\ldots\ldots\ldots\ldots\ldots \\ b_1 &= a_1 + ab_2, \\ b_0 &= a_0 + ab_1 = P_n(a). \end{aligned} \tag{3}$$

The method of finding the value of a polynomial by the formulas (3) (by the formula (2)) is known as *Horner's method.* The latter is realized with the aid of n multiplications and n additions, that is, by a total of $2n$ arithmetic operations. It is possible to prove that in the general case there is no method for computing an n-degree algebraic polynomial by performing less than $2n$ arithmetic operations. Horner's method can also be designed for computer calculations owing to the cyclicity of computations and necessity to preserve, in addition to the coefficients of the polynomial under computation, the values of the argument of only one intermediate quantity, namely, b_i or ab_i with a current $i = n, n - 1, \ldots, 1, 0$.

The manual evaluating of a polynomial by Horner's method is usually accompanied by the following table:

$$\begin{array}{cccccccc} & a_n & a_{n-1} & a_{n-2} & \dots & a_0 & & \lfloor a \\ + & & ab_n & ab_{n-1} & \dots & ab_1 & & \\ \hline b_n = & a_n & b_{n-1} & b_{n-2} & \dots & b_0 = P_n(a) & & \end{array}$$

Example. Evaluate for $x = -1.5$ the polynomial

$$P_5(x) = 1 - 4x + 3x^2 - x^3 + 2x^4 - x^5.$$

Solution.

$$\begin{array}{cccccccc} & -1 & 2 & -1 & 3 & -4 & 1 & \lfloor -1.5 \\ + & & 1.5 & -5.25 & 9.375 & -18.5625 & 33.84375 & \\ \hline & -1 & 3.5 & -6.25 & 12.375 & -22.5625 & 34.84375 = P_5(-1.5) & \end{array}$$

If a given polynomial is an even function, that is, $n = 2k$ and $P_{2k}(x) = a_0 + a_2x^2 + \dots + a_{2k}x^{2k}$, then it is readily computed by the formula

$$P_{2k}(x) = a_0 + x^2(a_2 + x^2(a_4 + \dots + x^2(a_{2k-4} + x^2(a_{2k-2} + x^2a_{2k}))\dots)), \quad (4)$$

and if the polynomial is an odd function, that is, $n = 2k + 1$ and $P_{2k+1}(x) = a_1x + a_3x^3 + \dots + a_{2k+1}x^{2k+1}$, then, prior to computations, it should be reduced to the form

$$P_{2k+1}(x) = x(a_1 + x^2(a_3 + \dots + x^2(a_{2k-3} + x^2(a_{2k-1} + x^2a_{2k+1}))\dots)). \quad (5)$$

We should like to conclude this section with the following remark: when using a computer to calculate the polynomials with very large coefficients with the aid of Horner's method (by the formula (2), (4) or (5)), a considerable loss of accuracy due to the subtraction of rounded-off numbers may occur. It is sometimes possible to avoid this loss by using recurrence formulas. A relevant example is considered at the close of Sec. 6.

Sec. 3. Taylor's Polynomials

Let there be given a function $f(x) \in C_{n+1}[a, b]$ (see Item 2 of Introduction). Let us recall: *Taylor's polynomial of degree n* of the function f at the point $x_0 \in [a, b]$ is defined as the polynomial

$$Q_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k. \tag{1}$$

Taylor's polynomial (1) possesses the property that at the point $x = x_0$ all of its derivatives till order n inclusive coincide with the corresponding derivatives of the function f, i.e.

$$Q_n^{(k)}(x_0) = f^{(k)}(x_0), \quad k = 0, 1, \ldots, n,$$

which is readily made sure by differentiating $Q_n(x)$. Thanks to this property, Taylor's polynomial approximates the function $f(x)$ in the neighbourhood of the point x_0 perfectly well. The error due to the replacement of the function f by its Taylor polynomial is expressed by the remainder of Taylor's formula*, that is,

$$f(x) - Q_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} (x - x_0)^{n+1}, \tag{2}$$

where $x \in [a, b]$ and ξ is a point lying exactly between x and x_0 for $x \neq x_0$. The given remainder is written in Lagrange's form which is convenient for obtaining error estimates. Since the derivative $f^{(n+1)}$ is, by the assumption, continuous on the interval $[a, b]$, it is bounded on this interval, that is,

$$M_{n+1} = \max_{[a, b]} |f^{(n+1)}(x)| < +\infty. \tag{3}$$

On the basis of (2), we have

$$|f(x) - Q_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |x - x_0|^{n+1}, \tag{4}$$

$$\max_{[a, b]} |f(x) - Q_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} l^{n+1}, \tag{5}$$

where $l = \max\{x_0 - a, b - x_0\}$.

---

* See Sec. 4.14 of the book *Differential and Integral Calculus* by Ya. S. Bugrov and S. M. Nikolsky (in English—Mir Publishers, Moscow, 1982).

The inequality (4) testifies to the fact that the error of approximating the function $f(x)$ by Taylor's polynomial (1) is the quantity* $O(|x - x_0|^{n+1})$, and the inequality (5) serves as an estimate of the maximum error on the entire interval $[a, b]$.

It is characteristic for the error of approximating a function by Taylor's polynomial to decrease rapidly as x approaches x_0 and to increase sharply at the end of the interval $[a, b]$ which is the most remote from the point x_0. In any case, it is so when $f^{(n+1)}(x) = \text{const} \neq 0$, that is, $f(x)$ is a polynomial of degree $n + 1$. Here $|f^{(n+1)}(x)| = M_{n+1}$, $x \in [a, b]$, and the inequalities (4) and (5) turn into equalities.

The accuracy of approximating the function f which is essentially nonuniform on the interval $[a, b]$ is one of the disadvantages of Taylor's polynomial, the other disadvantage consisting in that one has to find the derivatives of higher orders of the function f for constructing Taylor's polynomial. Nevertheless, Taylor's polynomials, in particular, the fragments of Taylor's series which are also Taylor's polynomials, are widely used in practice for approximating the functions whose higher derivatives are readily computed, while the remainder (the right-hand member of the inequality (5)) tends to zero as $n \to \infty$. Belonging to those functions are first of all the elementary functions $\sin x$, $\cos x$, e^x, $\ln(1 + x)$, and others.

Example. Approximate the function $f(x) = e^x$ by Taylor's polynomial on the interval $[0, 1]$ with absolute error not exceeding $\varepsilon = 10^{-5}$.

Solution. We choose $x_0 = 1/2$, that is, the midpoint of the interval $[0, 1]$, in order to minimize the quantity l entering into the right-hand member of the estimate (5). Then $f^{(k)}(x) = e^x$, $f^{(k)}(x_0) = e^{1/2}$, $M_{n+1} = e$, $l = 1/2$, and $Q_n(x) = e^{1/2} \sum_{k=0}^{n} \frac{1}{k!} \left(x - \frac{1}{2}\right)^k$. According to (5), we have

$$\max_{[0, 1]} |e^x - Q_n(x)| \leqslant r_n = \frac{e}{(n + 1)! 2^{n+1}}.$$

---

* For the definition of $O(|x - x_0|^{n+1})$ see Item 1 of Introduction (setting $|x - x_0| = h$).

For r_n we compile a table:

n	2	3	4	5	6
r_n	$5.7 \cdot 10^{-2}$	$7.1 \cdot 10^{-3}$	$7.1 \cdot 10^{-4}$	$5.9 \cdot 10^{-5}$	$4.3 \cdot 10^{-6}$

Thus, we should take $n = 6$.

Sec. 4. Lagrange's Interpolation Polynomial

Let there be known the values of some function f at $n + 1$ distinct points $x_0, x_1, \ldots, x_n$ which will be denoted as follows:

$$f_i = f(x_i), \quad i = 0, 1, \ldots, n.$$

For instance, these values are obtained experimentally or found with the aid of rather complicated computations.

There arises the problem of an approximate reconstruction of the function f at an arbitrary point x. Frequently, in order to solve this problem, we construct an algebraic polynomial $L_n(x)$ of degree n which attains the assigned values at the points x_i, that is,

$$L_n(x_i) = f_i, \quad i = 0, 1, \ldots, n, \tag{1}$$

which is called the *interpolation polynomial.* The points x_i, $i = 0, 1, \ldots, n$, are called the *interpolation points.*

For the sake of convenience, here and elsewhere by a polynomial of degree n we shall understand a polynomial of degree not higher than n. For instance, if $f_i = 0, i = 0, 1, \ldots, n$, then the interpolation polynomial $L_n(x) \equiv 0$ actually has a zero degree, but it will also be called the interpolation polynomial of degree n.

An approximate reconstruction of the function f by the formula

$$f(x) \approx L_n(x) \tag{2}$$

is called the *interpolation* of the function f (with the aid of an algebraic polynomial). If x is situated outside the minimum interval containing all the interpolation points $x_0, x_1, \ldots, x_n$, then the replacement of the function f by the formula (2) is also called *extrapolation.*

Let us first clarify the question of the existence and uniqueness of the interpolation polynomial and then investigate the interpolation error, that is, find out the difference between the left- and right-hand sides of the approximate equality (2).

Theorem 1. *There exists the unique interpolation polynomial of degree n satisfying the conditions* (1).

Proof. Let us establish the existence of the interpolation polynomial directly by writing out its expression. Let $n = 1$, then

$$L_1(x) = \frac{x - x_1}{x_0 - x_1} f_0 + \frac{x - x_0}{x_1 - x_0} f_1. \tag{3}$$

For $n = 2$

$$L_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2 \tag{4}$$

and, finally, in the general case for any natural n

$$L_n(x) = \sum_{i=0}^{n} p_{ni}(x) f_i, \tag{5}$$

where

$$p_{ni}(x) = \frac{(x - x_0)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}, \tag{6}$$

$$i = 0, 1, \ldots, n.$$

Indeed, the expression (3) represents a linear function, that is, a polynomial of first degree, and obviously, $L_1(x_0) = f_0$ and $L_1(x_1) = f_1$. Thus, the requirements (1) for $n = 1$ are fulfilled. Analogously, the formula (4) specifies a certain second-degree polynomial $L_2(x)$ which satisfies the conditions (1) for $n = 2$. For an arbitrary natural n the functions (6) expressed in the form of a common fraction whose numerator is the product of n linear multipliers and whose denominator is some nonzero number are algebraic polynomials of degree n. Consequently, the function (5) is also an algebraic polynomial of degree n, and since $p_{ni}(x_i) = 1$ and $p_{ni}(x_j) = 0$ for $j \neq i$, $0 \leqslant j \leqslant n$, the requirements (1) are fulfilled.

It remains to prove the uniqueness of the interpolation polynomial. Suppose that along with the interpolation polynomial (5) there is also some algebraic polynomial $\tilde{L}_n(x)$ of degree n satisfying the conditions

$$\tilde{L}_n(x_i) = f_i, \quad i = 0, 1, \ldots, n. \tag{7}$$

Then, according to (1) and (7),

$$\tilde{L}_n(x_i) - L_n(x_i) = 0, \quad i = 0, 1, \ldots, n. \tag{8}$$

If $\tilde{L}_n(x) - L_n(x) \not\equiv 0$, then this difference, being an algebraic polynomial of degree not higher than n, by virtue of the fundamental theorem of higher algebra, has at most n roots, which contradicts the equalities (8) whose number is equal to $n + 1$. Consequently, $\tilde{L}_n(x) \equiv L_n(x)$.

Theorem 1 has been proved completely.

The interpolation polynomial represented in the form (5) is called *Lagrange's interpolation polynomial,* and the functions (polynomials) (6)—the *Lagrangian coefficients.*

There are other forms of writing the interpolation polynomial. But, according to Theorem 1, the interpolation polynomial of degree n (speaking more exactly, of degree not higher than n) which satisfies the conditions (1) is unique. Another method of specifying the interpolation polynomial will be given in Sec. 9.

Remark 1. The actual degree of the interpolation polynomial (5) can be determined after removing the brackets and collecting like terms. But if at the interpolation points x_i the values of an algebraic polynomial $P_k(x)$ of degree $k \leqslant n$ are taken as f_i, then, by Theorem 1, automatically $L_n(x) \equiv P_k(x)$, since $P_k(x)$ is also a polynomial of degree not higher than n and satisfying the conditions (1). In particular, if $f_i = 1, i = 0, 1, \ldots, n$, then $L_n(x) \equiv P_0(x) \equiv 1$, wherefrom and from (5) there follows the identity $\sum_{i=0}^{n} p_{in}(x) \equiv 1$ which can serve as a check relation when computing the Lagrangian coefficients (6).

Remark 2. Since the interpolation polynomial (5) is linearly dependent on the values of the function f_i, the interpolation polynomial for the sum of two functions is equal to the sum of the interpolation polynomials for each term.

Example. Construct Lagrange's interpolation polynomial given the following data:

i	0	1	2	3
x_i	0	2	3	5
f_i	1	3	2	5

Solution. According to (5), for $n = 3$ we have

$$L_3(x) = \frac{(x-2)(x-3)(x-5)}{(0-2)(0-3)(0-5)} \cdot 1 + \frac{x(x-3)(x-5)}{2(2-3)(2-5)} \cdot 3 + \frac{x(x-2)(x-5)}{3(3-2)(3-5)} \cdot 2 + \frac{x(x-2)(x-3)}{5(5-2)(5-3)} \cdot 5 = 1 + \frac{62}{15}x - \frac{13}{6}x^2 + \frac{3}{10}x^3.$$

Interpolation Error. It is always possible to write the equality

$$f(x) = L_n(x) + R_n(x), \tag{9}$$

where $R_n(x)$ is the remainder, that is, the interpolation error. If nothing is known of the function f, except for its values f_i at the interpolation points, then no useful judgements can be made about the remainder $R_n(x)$. We shall obtain some expression for the remainder in the supposition that $f \in C_{n+1}[a, b]$, where $[a, b]$ is the interval containing all the interpolation points x_i, $i = 0, 1, \ldots, x_n$, and the point x.

We seek $R_n(x)$ in the following form:

$$R_n(x) = \omega_n(x) r_n(x), \tag{10}$$

where

$$\omega_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n), \tag{11}$$

and $r_n(x)$ is some function whose values at the interpolation points x_i may be assigned arbitrarily, since $R_n(x_i) = 0$ and $\omega_n(x_i) = 0, i = 0, 1, \ldots, n$.

Let us fix an arbitrary $x \in [a, b], x \neq x_i, i = 0, 1, \ldots, n$, and consider the following function of t:

$$\varphi(t) = L_n(t) + \omega_n(t) r_n(x) - f(t). \tag{12}$$

On the basis of (1) and (9)-(11) this function vanishes for $t = x_i, i = 0, 1, \ldots, n, t = x$, that is, at least at $n + 2$ points of the interval $[a, b]$ on which t varies. By Rolle's theorem, φ' (the first derivative with respect to t) vanishes at least at the $(n + 1)$st point of the interval (a, b), φ'' (the second derivative with respect to t) is equal to zero at least at n points of this interval, and so forth. Thus, there is at least one point $\xi \in (a, b)$ at which $\varphi^{(n+1)}(\xi) = 0$. Whence and from (12), bearing in mind that $L_n^{(n+1)}(\xi) = 0$, $\omega_n^{(n+1)}(\xi) = (n + 1)!$, we obtain

$$(n + 1)! r_n(x) - f^{(n+1)}(\xi) = 0.$$

Consequently,

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

and, in accordance with (9) and (10),

$$R_n(x) = \omega_n(x)\frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{13}$$

$$f(x) = L_n(x) + \omega_n(x)\frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{14}$$

where $\xi = \xi(x) \in (a, b)$ is some unknown point.

From the equality (14) there follows the estimate of the interpolation (in particular, extrapolation) error at the running point $x \in [a, b]$:

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega_n(x)| \tag{15}$$

and the estimate of the maximum interpolation error on the entire interval $[a, b]$:

$$\max_{[a, b]} |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} \max_{[a, b]} |\omega_n(x)|, \tag{16}$$

where M_{n+1} is the quantity (3.3).

Example. Estimate the error of approximating the function $f(x) = \sqrt{x}$ at the point $x = 116$ and on the entire interval $[a, b]$, where $a = 100$ and $b = 144$, with the aid of Lagrange's interpolation polynomial $L_2(x)$ of second degree constructed with the interpolation points $x_0 = 100$, $x_1 = 121$, and $x_2 = 144$.

Solution.

$$f'(x) = \frac{1}{2}x^{-1/2}, \quad f''(x) = -\frac{1}{4}x^{-3/2}, \quad f'''(x) = \frac{3}{8}x^{-5/2},$$

$$M_3 = \max_{[a, b]} |f'''(x)| = \frac{3}{8}100^{-5/2} = \frac{3}{8}10^{-5}.$$

On the basis of the inequality (15), we obtain

$$|\sqrt{116} - L_2(116)| \leqslant \frac{3}{8}10^{-5}\frac{1}{3!}|(116 - 100)(116 - 121)(116 - 144)|$$

$$= \frac{1}{16}10^{-5}\cdot 16\cdot 5\cdot 28 = 1.4\cdot 10^{-3}.$$

By virtue of the estimate (16),

$$\max_{[a,\,b]} |\sqrt{x} - L_2(x)| \leqslant \frac{10^{-5}}{16} \max_{[a,\,b]} |(x - 100)(x - 121)(x - 144)| \approx 2.5 \cdot 10^{-3}.$$

Sec. 5. Linear Interpolation

The interpolation by the formula (4.2) for $n = 1$, that is, with the aid of the linear function (4.3), is known as *linear interpolation*. Setting $h = x_1 - x_0$ and $q = (x - x_0)/h$, we can write the formula for linear interpolation in the following form:

$$f(x) \approx L_1(x) = L_1(x_0 + qh) = (1 - q)f_0 + qf_1. \tag{1}$$

The quantity q is called the *interpolation phase*. The latter changes within the limits from 0 to 1 as x runs through the values from x_0 to x_1.

Geometrically, linear interpolation means (Fig. 1) the replacement of the graph of the function on the interval $[x_0, x_1]$ by the chord joining the points (x_0, f_0) and (x_1, f_1).

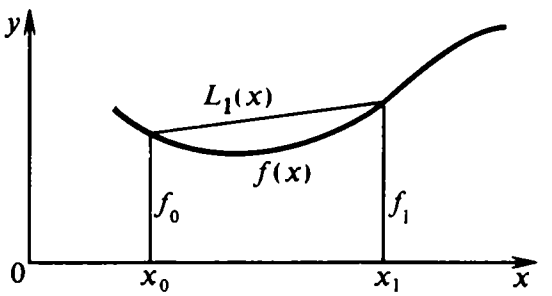


Fig. 1

Since, according to (4.11), $\omega_2(x) = (x - x_0)(x - x_1)$ and, consequently,

$$\max_{[x_0,\,x_1]} |\omega_2(x)| = \max_{[x_0,\,x_1]} |(x - x_0)(x - x_1)| = \frac{h^2}{4},$$

the estimate of the maximum error of the linear interpolation on the interval $[x_0, x_1]$, in accordance with (4.16), has the form

$$\max_{[x_0,\,x_1]} |f(x) - L_1(x)| \leqslant h^2 \frac{M_2}{8}, \tag{2}$$

where $M_2 = \max_{[x_0,\,x_1]} |f''(x)|$.

Frequently, a table is given containing a great number of values of some function f with constant step h of argument variation. Then for a given x two nearest to it interpolation points are chosen. The left-hand point is taken for x_0 and the right-hand point for x_1, and linear interpolation by the formula (1) is realized. The interpolation error is estimated by the formula (2).

Example. Given a table of the function $\sin x$ with a step of $1°$. It is required to estimate the error of linear interpolation.

Solution. In radian measure, the step h of the given table is equal to $\pi/180$. Since $M_2 = \max |(\sin x)''| \leqslant 1$, the error of linear interpolation, according to (2), does not exceed the quantity

$$\rho_1 = \frac{\pi^2}{180^2} \cdot \frac{1}{8} < 0.4 \cdot 10^{-4}.$$

The quantity ρ_1 serves as the estimate of the error due to the replacement on the interval of length h of the function $\sin x$ by the interpolation polynomial $L_1(x)$ in the supposition that the tabular values are exact and the computations by the formula (1) are carried out without errors.

Let us now assume that the tabular values of the function $\sin x$ are rounded off to four places after the decimal point. Then the limiting absolute error of the rounded-off tabular values is equal to $\bar{\Delta}(f) = 0.5 \cdot 10^{-4}$. If rounded-off values are substituted into the formula (1) for the exact values f_0 and f_1 of the function $f(x) = \sin x$, then in $L_1(x)$ there occurs an error which can be estimated by the quantity

$$\bar{\Delta}(L_1) = (1 - q)\bar{\Delta}(f) + q\bar{\Delta}(f) = \bar{\Delta}(f) = 0.5 \cdot 10^{-4},$$

since $0 \leqslant q \leqslant 1$.

If the value of the interpolation polynomial $L_1(x)$ thus found is finally rounded off to four decimal places, then an additional error appears which is bounded in modulus by the quantity $\bar{\Delta}_1(L_1) = 0.5 \cdot 10^{-4}$.

Thus, the total error of the linear interpolation of the function $\sin x$ for the rounded-off tabular values and rounding off the final result to four places after the decimal point is estimated by the quantity

$$\rho_1 + \bar{\Delta}(L_1) + \bar{\Delta}_1(L_1) < 1.4 \cdot 10^{-4}.$$

Sec. 6. Minimizing the Interpolation Error Estimate. Chebyshev Polynomials

Let there be given a certain function $f \in C_{n+1}[a, b]$. There arises a question as to how to choose the points $x_0, x_1, \ldots, x_n$ of the interpolation polynomial (4.5) on the interval $[a, b]$ so that the maximum interpolation

error for the function f be minimal on this interval. This problem is complicated and is successfully solvable only for particular functions f. We shall confine ourselves to solving a simpler problem, namely, to finding such an arrangement of the interpolation points x_i, $i = 0, 1, \ldots, n$, on the interval $[a, b]$ for which the quantity $\max\limits_{[a, b]} |\omega_n(x)|$ is minimal and, thus, the right-hand side of the error estimate (4.16) is also minimal.

For the sake of simplicity, let us first consider the case of the standard interval $[-1, 1]$. For this purpose, Chebyshev polynomials are required.

Chebyshev Polynomials. *Chebyshev polynomials* $T_n(x)$, $n \geqslant 0$, on the interval $[-1, 1]$ are defined by the formula

$$T_n(x) = \cos (n \arccos x). \tag{1}$$

In particular, for $n = 0$ and 1 we have

$$T_0(x) = \cos (0 \cdot \arccos x) = 1, \tag{2}$$

$$T_1(x) = \cos (\arccos x) = x. \tag{3}$$

Further, from the identity

$$\cos (n + 1)\varphi = 2 \cos \varphi \cos n\varphi - \cos (n - 1)\varphi,$$

setting $\varphi = \arccos x$, in accordance with (1), we get

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \tag{4}$$

where $n = 1, 2, \ldots$. Thus, $T_n(x)$ is actually an algebraic polynomial of degree $n \geqslant 0$.

Setting $T_0(x) = 1$ and $T_1(x) = x$ throughout the x-axis and extending the recurrence formula (4) to the entire x-axis, we find successively by the formula (4):

$$T_2(x) = 2x^2 - 1, \qquad T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1, \qquad T_5(x) = 16x^5 - 20x^3 + 5x, \ldots .$$

Properties of Chebyshev Polynomials.

Property 1. *For an even (odd) n the polynomial* $T_n(x)$ *contains only even (odd) powers of x, i.e. it is an even (odd) function.*

Property 1 follows readily from the formulas (2)-(4).

Property 2. *The leading coefficient of the polynomial* $T_n(x)$ *for* $n \geqslant 1$ *is equal to* 2^{n-1}.

This property follows from the formulas (2)-(4).

Property 3. *The polynomial $T_n(x)$ has n real roots on the interval* $(-1, 1)$ *which are expressed by the formula*

$$x_i = \cos \frac{(2i + 1)\pi}{2n}, \quad i = 0, 1, \ldots, n - 1.$$

Indeed, $T_n(x_i) = \cos (n \arccos x_i) = \cos \dfrac{(2i + 1)\pi}{2} = 0$, $i = 0, 1, \ldots, n - 1$.

Property 4. $\max\limits_{[-1, 1]} |T_n(x)| = 1$, *and*

$$T_n(x_m) = (-1)^m, \tag{5}$$

where $x_m = \cos (m\pi/n)$, $m = 0, 1, \ldots, n$.

Indeed, according to (1), $T_n(x_m) = \cos m\pi = (-1)^m$ and $|T_n(x)| \leqslant 1, x \in [-1, 1]$. Consequently, Property 4 holds true.

Property 5. *Among all n-degree polynomials with leading coefficient equal to unity the polynomial*

$$\bar{T}_n(x) = 2^{1-n} T_n(x), \quad n \geqslant 1, \tag{6}$$

has the least maximum-modulus value on the interval $[-1, 1]$, *that is, there is no polynomial* $\bar{P}_n(x)$ *of degree n with leading coefficient* 1 *such that*

$$\max_{[-1, 1]} |\bar{P}_n(x)| < \max_{[-1, 1]} |\bar{T}_n(x)| = 2^{1-n}. \tag{7}$$

Assume the contrary: there is a polynomial $\bar{P}_n(x) = a_0 + \ldots + a_{n-1}x^{n-1} + x^n$ which satisfies the inequality (7). Then since, by Property 2, the leading coefficient of the polynomial $\bar{T}_n(x)$ is also equal to 1, the difference $\bar{T}_n(x) - \bar{P}_n(x)$ is an algebraic polynomial of degree not higher than $n - 1$, and, by virtue of (7), $\bar{T}_n(x) - \bar{P}_n(x) \not\equiv 0$. Besides, at $n + 1$ points $x_m = \cos (m\pi/n)$, $m = 0, 1, \ldots, n$, on the basis of (5)-(7), this difference attains nonzero values with alternating signs. This means that the algebraic polynomial $\bar{T}_n(x) - \bar{P}_n(x)$ of degree lower than n vanishes at least at n points, which is impossible.

Property 5 has been proved.

Remark 1. It is possible to prove that if $\bar{P}_n(x) = a_0 + \ldots + a_{n-1}x^{n-1} + x^n$, where $n \geqslant 1$, $\max\limits_{[-1, 1]} |\bar{P}_n(x)| = 2^{1-n}$, then $\bar{P}_n(x) = 2^{1-n} T_n(x) = \bar{T}_n(x)$.

Owing to Property 5, the Chebyshev polynomials $T_n(x)$ are called *polynomials least deviating from zero.*

Interpolation Points Minimizing the Interpolation Error Estimate. Let us take the roots of the Chebyshev polynomial $T_{n+1}(x)$, that is, the points

$$x_i = \cos \frac{(2i + 1)\pi}{2n + 2}, \quad i = 0, 1, \ldots, n, \tag{8}$$

on the interval $[-1, 1]$ as the interpolation points. Then the polynomial (4.11) whose leading coefficient is equal to unity will be proportional to the polynomial $T_{n+1}(x)$ and, by virtue of Property 2 of Chebyshev polynomials, it will be expressed in terms of $T_{n+1}(x)$ in the following way:

$$\omega_n(x) = 2^{-n} T_{n+1}(x).$$

And in accordance with Property 4, the interpolation error estimate (4.16) will take the form

$$\max_{[-1, 1]} |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n + 1)!2^n}, \tag{9}$$

where $M_{n+1} = \max\limits_{[-1, 1]} |f^{(n+1)}(x)|$.

By virtue of Property 5 of Chebyshev polynomials, it is impossible to improve the estimate (4.16) on the interval $[-1, 1]$ in comparison with the estimate (9) by a different choice of interpolation points. Moreover, according to Property 5 and Remark 1, for any choice of the interpolation points not coinciding with (8) the appropriate estimate of the maximum interpolation error on the interval $[-1, 1]$ will be worse, that is, *the interpolation points* (8) *are optimal for the error estimate* (4.16) *on the interval* $[-1, 1]$.

In the case of interpolation performed on an arbitrary interval $[a, b]$ the latter is transformed into the interval $[-1, 1]$ by the following linear replacement of the independent variable:

$$x = \frac{1}{2}((b - a)t + b + a), \quad t = \frac{1}{b - a}(2x - b - a). \tag{10}$$

Here to the roots of the Chebyshev polynomial $T_{n+1}(t)$ there correspond the points

$$x_i = \frac{1}{2}\left((b - a)\cos\frac{(2i + 1)\pi}{2n + 2} + b + a\right), \quad i = 0, 1, \ldots, n, \tag{11}$$

of the interval $[a, b]$ which are the optimal interpolation points for esti-

mating the interpolation error on this interval. According to (8), (10),and (11), we shall have

$$\omega_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$$
$$= \frac{(b-a)^{n+1}}{2^{n+1}}\left(t - \cos\frac{\pi}{2n+2}\right)\left(t - \cos\frac{3\pi}{2n+2}\right)\ldots$$
$$\ldots\left(t - \cos\frac{(2n+1)\pi}{2n+2}\right) = \frac{(b-a)^{n+1}}{2^{n+1}}\bar{T}_{n+1}(t).$$

Hence, with (6) and Property 4 in mind, we get

$$\max_{[a,\,b]}|\omega_n(x)| = \frac{(b-a)^{n+1}}{2^{n+1}}\max_{[-1,\,1]}|\bar{T}_{n+1}(t)| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Consequently, with the interpolation points (11) the interpolation error estimate (4.16) takes the form

$$\max_{[a,\,b]}|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}, \tag{12}$$

where $M_{n+1} = \max\limits_{[a,\,b]}|f^{(n+1)}(x)|$.

Let us now compare the methods for approximating the function $f \in C_{n+1}[a, b]$ by Taylor's polynomial $Q_n(x)$ and by the interpolation polynomial $L_n(x)$ with the interpolation points (11). When constructing Taylor's polynomial (3.1), it is advisable to take the point x_0 in the middle of the interval $[a, b]$, that is, $x_0 = (a + b)/2$. Then, in accordance with (3.5), the maximum error of Taylor's polynomial will be estimated as follows:

$$\max_{[a,\,b]}|f(x) - Q_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}\frac{(b-a)^{n+1}}{2^{n+1}}. \tag{13}$$

Thus, the error estimate of Taylor's polynomial is 2^n times the error estimate (12) of Lagrange's interpolation polynomial $L_n(x)$ with the optimal interpolation points (11). If the derivative $f^{(n+1)}(x)$ changes on the interval $[a, b]$ but relatively little, then both estimates (12) and (13) do not overstate the maximum error to a considerable extent, and if $f^{(n+1)}(x) = \text{const}$, then either of the inequalities (12) and (13) turns into an equality. The error of the interpolation polynomial is distributed more uniformly on the interval $[a, b]$ than in Taylor's polynomial which has an essentially nonuniform error (see Sec. 3). Therefore, not only the error estimate but also the actual maximum error of the interpolation

polynomial are usually less on the entire interval $[a, b]$ than in Taylor's polynomial. In addition, to construct the interpolation polynomial, it is not required to compute the derivatives of the function f, we need only its values.

Thus, preference should be given to the interpolation polynomial with the optimal interpolation points (11).

Example. Let there be given the function $f(x) = e^x$ on the interval $[-1, 1]$. Its Taylor polynomial $Q_5(x)$ at the point $x_0 = 0$ has the form $Q_5(x) = \sum_{k=0}^{5} \frac{x^k}{k!}$. Since $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, we have

$$\max_{[-1, 1]} |e^x - Q_5(x)| = \max_{[-1, 1]} \left| \sum_{k=6}^{\infty} \frac{x^k}{k!} \right| > \frac{1}{6!} + \frac{1}{7!} > 0.15 \cdot 10^{-2}.$$

As to Lagrange's interpolation polynomial $L_5(x)$ with the interpolation points (8) for $n = 5$, we can state that since, according to (9), $M_6 = \max_{[-1, 1]} \left| \frac{d^6}{dx^6} e^x \right| = e$, the following inequality is fulfilled for it:

$$\max_{[-1, 1]} |e^x - L_5(x)| \leqslant \frac{e}{6! 2^5} < 0.12 \cdot 10^{-3}.$$

Thus, in the discussed example the upper bound of the interpolation polynomial $L_5(x)$ is considerably less than the lower bound of the maximum error of Taylor's polynomial $Q_5(x)$. This example confirms the above findings.

In conclusion, let us dwell on evaluating Chebyshev polynomials. According to Properties 2 and 4, on the one hand, the values of the polynomials $T_n(x)$ for any n are bounded on the interval $[-1, 1]$ by unity (in magnitude). On the other hand, there occurs a rapid growth of the coefficients in these polynomials with an increase in n. For instance, the leading coefficient of the polynomial $T_n(x)$ is equal to 2^{n-1}, in particular, for $n = 21$ we have $2^{n-1} > 10^6$. This leads to the fact that using Horner's method (2.2) we evaluate the polynomial $T_n(x)$ for $x \in [-1, 1]$ on a computer with a limited number of digit places, this evaluation being realized for large n's with a considerable loss in accuracy. A loss in accuracy is caused by the necessity of subtracting close rounded-off numbers large in absolute value, since the final result does not exceed unity in absolute value.

The recurrence formula (4) enables us to avoid this difficulty. Using directly the formulas (2) and (3), we first find $T_0(x)$ and $T_1(x)$ for the given $x \in [-1, 1]$ and then via the formula (4) successively compute the values of $T_n(x)$ for increasing n's, the given n included. It is possible to prove that in this case the total round-off error in the value of $T_n(x)$ is increased by no more than $O(n^2)$ times as compared with the maximum round-off error in one-time computation by the formula (4). In any case, large quantities are not involved in computing by the formula (4) for $x \in [-1, 1]$, since $|T_n(x)| \leqslant 1$ and $|T_{n-1}(x)| \leqslant 1$ for any natural n.

And in computing the polynomials $T_n(x)$ by Horner's method, the effect of round-off errors with an increase in n is more pronounced than by the formula (4).

Sec. 7. Interpolation with Equally Spaced Points

In the preceding section, an optimal distribution of interpolation points was found which ensures the minimum error estimate on the entire interval. The optimal distribution of interpolation points is nonuniform: they are concentrated towards the ends of an interval and are spaced at greater distances in the middle (see (6.8) and (6.11)). This proves to be inconvenient for practical purposes. A nonuniform distribution of interpolation points may be specified, say, if a certain interpolation polynomial is constructed once and for the entire interval. But if the interval $[a, b]$ is lengthy and a high accuracy of approximation to a function is required, then one interpolation polynomial of acceptable degree even with an optimal distribution of interpolation points usually fails to ensure the assigned accuracy of interpolation on the entire interval. In such a case, we often use the table of the function values at equally spaced interpolation points (that is, arranged at a constant step), the number of such points being sufficiently large. When the value of the argument x is specified, we choose several nearest to x interpolation points to be used for constructing an interpolation polynomial, usually of a low degree, and carry out the interpolation. Given below is the relation between the accuracy of interpolation and the step at which the interpolation points are spaced.

Thus, let $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$, be interpolation points, $h > 0$ the step, $f_i = f(x_i)$ the given values of the function $f \in C_{n+1}[a, b]$, and $[x_0, x_n] \subset [a, b]$.

Let us introduce the dimensionless independent variable

$$q = \frac{x - x_0}{h} \quad (x = x_0 + qh). \tag{1}$$

Then

$$q = q_i = \frac{x_i - x_0}{h} = \frac{x_0 + ih - x_0}{h} = i \tag{2}$$

will correspond to the interpolation point x_i, and furthermore, the following relationships will be fulfilled:

$$x - x_j = h(q - j), \quad x_i - x_j = h(i - j). \tag{3}$$

Here, the interpolation polynomial (4.3) which corresponds to the case $n = 1$ will be written as indicated in the formula (5.1), and the second-degree interpolation polynomial (4.4) takes the form

$$L_2(x) = L_2(x_0 + qh) = \frac{(q-1)(q-2)}{2} f_0 - q(q-2)f_1 + \frac{q(q-1)}{2} f_2.$$

In the general case, with (1)-(3) and (4.6) taken into account, the interpolation polynomial (4.5) will take the following form:

$$L_n(x) = L_n(x_0 + qh) = \sum_{i=0}^{n} \bar{p}_{ni}(q) f_i, \tag{4}$$

where

$$\bar{p}_{ni}(q) = (-1)^{n-i} \frac{q(q-1)\ldots(q-i+1)(q-i-1)\ldots(q-n)}{i!(n-i)!}. \tag{5}$$

Since, according to (3),

$$\omega_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n) = h^{n+1} \bar{\omega}_n(q),$$

where

$$\bar{\omega}_n(q) = q(q-1)\ldots(q-n), \tag{6}$$

the remainder (4.13) of the interpolation polynomial (4) can be represented in the following form:

$$R_n(x) = R_n(x_0 + qh) = h^{n+1} \bar{\omega}_n(q) \frac{f^{(n+1)}(\xi).}{(n+1)!}. \tag{7}$$

Here, $f^{(n+1)}$ is the derivative with respect to x, and the intermediate point $\xi = \xi(x)$ is the same as in (4.13) and (4.14).

Note that, according to (1), the change of the variable q on the interval $[0, n]$ corresponds to the change of the variable x on the interval $[x_0, x_n]$, where $x_n = x_0 + nh$. Therefore the estimate of the maximum interpolation error on the interval $[x_0, x_n]$, with (4.9) and (7) taken into account, may be written in the following form:

$$\max_{[x_0, x_n]} |f(x) - L_n(x)| \leqslant h^{n+1} \frac{M_{n+1}^h}{(n+1)!} \Omega_n, \tag{8}$$

where $\Omega_n = \max_{[0, n]} |\bar{\omega}_n(q)|$,

$$M_{n+1}^h = \max_{[x_0, x_n]} |f^{(n+1)}(x)|. \tag{9}$$

The quantity Ω_n is independent of h. It can be computed or estimated beforehand. In particular,

$$\Omega_1 = \frac{1}{4}, \quad \Omega_2 = \frac{2}{3\sqrt{3}}, \quad \Omega_3 = 1, \quad \Omega_4 < 3.7, \quad \Omega_5 < 17. \tag{10}$$

For small h the quantity M_{n+1}^h differs but little from $|f^{(n+1)}(x_0)|$, since $x_n \to x_0$ as $h \to 0$, and the derivative $f^{(n+1)}$, by the supposition, is continuous in some neighbourhood of the point x_0. In any case, if $[x_0, x_n] \subset [a, b]$, then

$$M_{n+1}^h \leqslant M_{n+1}, \tag{11}$$

where $M_{n+1} = \max_{[a, b]} |f^{(n+1)}(x)|$.

Thus, according to (8) and (11), the maximum interpolation error on the interval $[x_0, x_n]$, that is, $\max_{[x_0, x_n]} |f(x) - L_n(x)|$ is the quantity $O(h^{n+1})$. Note, for instance, that, since $M_{n+1}^{h/2} \leqslant M_{n+1}^h$, when halving the step h, the right-hand member of the estimate (8) decreases at least 2^{n+1} times.

Proceeding from the strong estimate obtained by the inequality (8) in which M_{n+1} is substituted for the quantity M_{n+1}^h, in accordance with (11), we choose the step h of the tabular values of the function f on the interval $[a, b]$ in order to ensure the assigned accuracy of interpolation. Here, there is a possibility to vary within some limits the degree n of an interpolation polynomial. If the function f is sufficiently smooth, then an increase in n at first usually leads to an increase in the admissible h, but, on the other hand, makes the interpolation more complicated and strengthens the effect of inherent errors of the table of the values. In practice, interpolation with $n \geqslant 5$ is used rarely.

Remark 1. For a given x it is advisable to choose the interpolation points $x_0, x_1, \ldots, x_n$ spaced at the step h from all the points of the given table of the function so that the point x turns out to be as near as possible to the middle of the interval $[x_0, x_n]$. This is connected with the fact that the fluctuations of the function (4.11) with respect to the function (6) near the middle of the indicated interval are smaller than at its ends. Shown in Fig. 2 is the graph of the function (4.11) for $n = 4$.

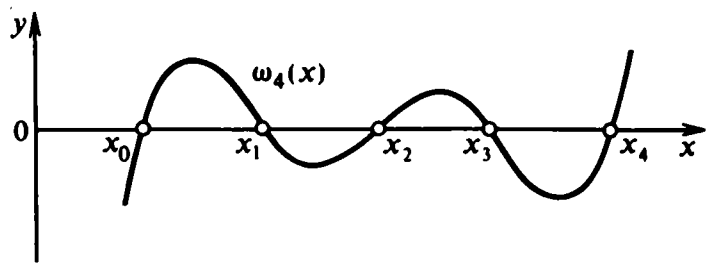


Fig. 2

If the point x at which interpolation is performed satisfies the inequality

$$\left|x - \frac{x_0 + x_n}{2}\right| \leqslant \frac{h}{2}, \tag{12}$$

then for the interpolation error the following estimate is true:

$$|f(x) - L_n(x)| \leqslant h^{n+1} \frac{M_{n+1}^h}{(n+1)!} \Omega_n^0, \tag{13}$$

where $\Omega_n^0 = \max_{|q - n/2| \leqslant 1/2} |\bar{\omega}_n(q)|$. In particular,

$$\Omega_1^0 = \frac{1}{4}, \quad \Omega_2^0 = \frac{3}{8}, \quad \Omega_3^0 = \frac{9}{16}, \quad \Omega_4^0 = \frac{45}{32}, \quad \Omega_5^0 = \frac{225}{64}. \tag{14}$$

Comparing the quantities (14) and (10), we see that the interpolation error estimate (13) is more exact than (8) for x's satisfying the condition (12).

Sec. 8. Finite and Divided Differences

Finite Differences. Let $x_k = x_0 + kh$, where k is an integer, h is the step, or tabulated interval, $h > 0$, and $f_k = f(x_k)$. The quantity

$$\Delta f_k = f(x_k + h) - f(x_k) = f(x_{k+1}) - f(x_k) = f_{k+1} - f_k \tag{1}$$

is called the *finite difference of the first order* of the function f at the point x_k (with step h), and

$$\Delta^2 f_k = \Delta(\Delta f_k) = \Delta f_{k+1} - \Delta f_k$$
$$= (f_{k+2} - f_{k+1}) - (f_{k+1} - f_k) = f_k - 2f_{k+1} + f_{k+2} \quad (2)$$

is the *finite difference of the second order* at the point x_k. In general, the nth-*order finite difference* of the function f at the point x_k is determined by the recurrence formula

$$\Delta^n f_k = \Delta(\Delta^{n-1} f_k) = \Delta^{n-1} f_{k+1} - \Delta^{n-1} f_k, \quad (3)$$

where $n \geqslant 1$ and $\Delta^0 f_k = f_k$.

When carrying out computations, it is convenient to write finite differences in tabular form (Table 1).

Table 1

x_0	f_0			
		Δf_0		
x_1	f_1		$\Delta^2 f_0$	
		Δf_1		$\Delta^3 f_0$
x_2	f_2		$\Delta^2 f_1$	
		Δf_2		$\Delta^3 f_1$
x_3	f_3		$\Delta^2 f_2$	
		Δf_3		
x_4	f_4			

Lemma 1. *If* $f \in C_n[x_k, x_{k+n}]$, *then there is a point* $\eta \in (x_k, x_{k+n})$ *such that*

$$\Delta^n f_k = h^n f^{(n)}(\eta). \quad (4)$$

By virtue of (1), the formula (4) for $n = 1$ coincides with Lagrange's formula of finite increments. Let us prove it for $n = 2$. Let $\varphi(x) = f(x + h) - f(x)$. Then according to (3),

$$\Delta^2 f_k = \varphi(x_k + h) - \varphi(x_k)$$

and, consequently, by Lagrange's formula of finite increments

$$\Delta^2 f_k = h\varphi'(\eta_1), \quad (5)$$

where $\eta_1 \in (x_k, x_k + h)$ is some point. But

$$\varphi'(\eta_1) = f'(\eta_1 + h) - f'(\eta_1).$$

Applying Lagrange's formula of finite increments to the function f' once again, we obtain

$$\varphi'(\eta_1) = hf''(\eta), \tag{6}$$

where $\eta \in (\eta_1, \eta_1 + h) \subset (x_k, x_{k+2})$ is some point. From (5) and (6) there follows the statement of Lemma 1 for $n = 2$. For $n > 2$ Lemma 1 is proved in a similar way.

Corollary of Lemma 1. *The finite difference of order n of an algebraic polynomial of degree n is constant, i.e. is independent of k, while the finite differences of any higher order are equal to zero.*

Let us dwell on one of the practical applications of finite differences. According to Lemma 1, if $f \in C_{n+1}[x_0, x_{n+1}]$, then the quantity $\Delta^{n+1} f_0/h^{n+1}$ which can be computed from the tabular values of the function f with the aid of the formula (3) is equal to the value of the derivative $f^{(n+1)}$ at some point $\eta \in (x_0, x_{n+1})$, where $x_{n+1} = x_0 + (n+1)h$. Therefore, if h is small, then the number $|\Delta^{n+1} f_0/h^{n+1}|$ may be approximately taken for the quantity (7.9) and used in the estimate (7.8) of the interpolation error. Such a nonstrict estimate of the error is used if the computation of the derivative $f^{(n+1)}(x)$ is rather complicated or, in general, if there are only the tabular values of $n + 1$ times differentiable function at our disposal.

Divided Differences. Let now $x_0, x_1, \ldots, x_k, \ldots$ be arbitrary points (nodes) on the x-axis, and $x_i \neq x_j$ for $i \neq j$. The values $f(x_0), f(x_1), \ldots$ of the function f at these points are called the *divided differences of zeroth order*. The number

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{7}$$

is called the *divided difference of the first order* of the function f. Obviously,

$$f(x_0; x_1) = f(x_1; x_0) = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}, \tag{8}$$

that is, the divided difference of first order is a symmetric function of the arguments x_0 and x_1.

The *divided difference of order n* is defined via the divided differences of order $n - 1$ by the recurrence formula

$$f(x_0; x_1; \ldots; x_n) = \frac{f(x_1; x_2; \ldots; x_n) - f(x_0; x_1; \ldots; x_{n-1})}{x_n - x_0}. \tag{9}$$

For carrying out computations, the divided differences are written in the form of array (Table 2).

Table 2

x_0	$f(x_0)$			
		$f(x_0; x_1)$		
x_1	$f(x_1)$		$f(x_0; x_1; x_2)$	
		$f(x_1; x_2)$		$f(x_0; x_1; x_2; x_3)$
x_2	$f(x_2)$		$f(x_1; x_2; x_3)$	
		$f(x_2; x_3)$		$f(x_1; x_2; x_3; x_4)$
x_3	$f(x_3)$		$f(x_2; x_3; x_4)$	
		$f(x_3; x_4)$		
x_4	$f(x_4)$			

Lemma 2. *The nth-order divided difference is expressed in terms of the nodal values of the function by the formula*

$$f(x_0; x_1; \ldots; x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}, \tag{10}$$

i.e. it is a symmetric function of its arguments.

For $n = 1$ this statement follows from the equality (8). For $n = 2$, in accordance with (9), we have

$$f(x_0; x_1; x_2) = \frac{1}{x_2 - x_0}\left(\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}\right)$$

$$= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}. \tag{11}$$

The equality (10) has been proved for $n = 2$. The proof of the equality (10) for an arbitrary n is left to the reader (*hint:* use the method of induction).

Thus, according to Lemma 2, the value of the nth-order divided difference is independent of the numeration of $n + 1$ points by which it is constructed. There is a total of $(n + 1)!$ different variants of their numbering by integers from 0 to n.

Lemma 3. *If $x_k = x_0 + kh$, $k = 0, 1, \ldots$, that is, if the nodal points are spaced at a constant step $h > 0$, then there is the following relation between the nth-order divided difference and the nth-order finite dif-*

ference:

$$f(x_0; x_1; \ldots; x_n) = \frac{\Delta^n f_0}{n! h^n}. \tag{12}$$

Proof. For $n = 1$ the equality (12) follows from (1) and (7). When finding each subsequent finite difference, according to (3), we simply subtract the previous differences, and when computing each subsequent divided difference, in accordance with the formula (9), in addition to subtraction, we divide it by the quantity $x_n - x_0 = nh$. That is why the quantity $n!h^n$ appears in the denominator of the right-hand member of the equality (12).

Lemma 4. *Let* $[\alpha, \beta]$ *be the minimum interval containing the points* x_0, $x_1, \ldots, x_n$, $f \in C_n[\alpha, \beta]$. *Then there is a point* $\eta \in (\alpha, \beta)$ *such that*

$$f(x_0; x_1; \ldots; x_n) = \frac{f^{(n)}(\eta)}{n!}. \tag{13}$$

For the points spaced at a constant step, the equality (13) follows from Lemmas 1 and 3. We are not going to prove Lemma 4 for the general case.

Corollary of Lemma 4. *The nth-order divided difference of an nth-degree algebraic polynomial takes on a constant value independent of the choice of the points* $x_0, x_1, \ldots, x_n$, *and the divided differences of higher orders are equal to zero.*

Finite and divided differences have various applications. In Sec. 9 they are used for constructing Newton's interpolation polynomial.

Sec. 9. Newton's Interpolation Polynomial

Let $x_0, x_1, \ldots, x_n$ be arbitrary pairwise noncoinciding interpolation points at which the values of the function f are known.

Lemma 1. *The nth-degree algebraic polynomial*

$$\begin{aligned} l_n(x) = f(x_0) &+ (x - x_0) f(x_0; x_1) \\ &+ (x - x_0)(x - x_1) f(x_0; x_1; x_2) + \\ &\ldots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f(x_0; x_1; \ldots; x_n) \end{aligned} \tag{1}$$

is an interpolation one, that is,

$$l_n(x_i) = f(x_i), \quad i = 0, 1, \ldots, n. \tag{2}$$

Let us first of all note that since the divided differences $f(x_0; x_1)$, $f(x_0; x_1; x_2)$, ..., $f(x_0; x_1; \ldots; x_n)$ are quite definite numbers (see. Sec. 8), the function (1) is actually an algebraic polynomial of degree n.

Let us prove the equalities (2) for $n = 2$. We have

$$l_2(x) = f(x_0) + (x - x_0)f(x_0; x_1) + (x - x_0)(x - x_1)f(x_0; x_1; x_2). \quad (3)$$

Obviously, $l_2(x_0) = f(x_0)$. Further, according to (3) and (8.7),

$$l_2(x_1) = f(x_0) + (x_1 - x_0)f(x_0; x_1)$$

$$= f(x_0) + (x_1 - x_0)\frac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_1).$$

Finally, taking into account (3), (8.7),and (8.11), we obtain

$$l_2(x_2) = f(x_0) + (x_2 - x_0)f(x_0; x_1)$$

$$+ (x_2 - x_0)(x_2 - x_1)f(x_0; x_1; x_2)$$

$$= f(x_0) + \frac{x_2 - x_0}{x_1 - x_0}(f(x_1) - f(x_0)) + \frac{(x_2 - x_0)(x_2 - x_1)}{(x_0 - x_1)(x_0 - x_2)}f(x_0)$$

$$+ \frac{(x_2 - x_0)(x_2 - x_1)}{(x_1 - x_0)(x_1 - x_2)}f(x_1) + f(x_2) = f(x_2).$$

For $n = 1$ the equalities (2) are proved in a similar way, and for an arbitrary natural $n > 2$ they are proved by induction.

Remark 1. The polynomial (3) is known as *Newton's interpolation polynomial for unequal intervals.* According to Theorem 4.1, it coincides identically with Lagrange's interpolation polynomial (4.5), that is, $l_n(x) \equiv L_n(x)$.

Thus, we have two different notations for the interpolation polynomial. Newton's and Lagrange's interpolation polynomials have the same remainder, that is, everywhere in the equalities (4.9), (4.14) and the inequalities (4.15), (4.16) we may replace $L_n(x)$ by $l_n(x)$.

In Lagrange's interpolation polynomial (4.5) we see its explicit dependence on each value of the function f_i, $i = 0, 1, \ldots, n$. In many cases this proves to be useful. But with a change in n, Lagrange's interpolation polynomial must be reconstructed, this being its disadvantage.

Newton's interpolation polynomial (1) is expressed not in terms of the values of the function f but via its divided differences. With a change in the degree n of Newton's interpolation polynomial, the only thing required is to add or reject the appropriate number of standard terms. In practice, this is convenient.

The Case of Equally Spaced Interpolation Points. Let $x_k = x_0 + kh$, $h > 0$, $k = 0, 1, \ldots, n$, and $f_k = f(x_k)$. Then, taking into consideration the relation between the divided and finite differences (8.12) and introducing the dimensionless variable q (7.1), we may rewrite the interpolation polynomial (1) in the following form:

$$l_n(x) = l_n(x_0 + qh) = f_0 + q\frac{\Delta f_0}{1!}$$

$$+ q(q-1)\frac{\Delta^2 f_0}{2!} + \ldots + q(q-1)\ldots(q-n+1)\frac{\Delta^n f_0}{n!}. \quad (4)$$

It is called *Newton's interpolation polynomial for forward interpolation.* In it, the reference point q is situated at the leftmost interpolation point x_0, and the finite differences used are arranged in the table of differences from f_0 to the right downwards (see Table 8.1).

The interpolation polynomial (4) may be conveniently used for interpolation at the beginning of the table and for extrapolation to the left of the point x_0, that is, for $q < 0$.

Consider an example of interpolation by the formula (4). Let there be given a table of the values of the function $f(x) = \sin x$ and its finite differences:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
5°	0.087156			
		34,713		
7°	0.121869		−148	
		34,565		−42
9°	0.156434		−190	
		34,375		−43
11°	0.190809		−233	
		34,142		−41
13°	0.224951		−274	
		33,868		
15°	0.258819			

For the sake of simplicity, the finite differences of a function are customarily written without an explicit indication of the position of the decimal point.

Suppose that it is required to find sin 6°. It is seen from the table that the third differences are close to a constant. According to the corollary of Lemma 8.1, this testifies to the fact that the function $f(x) = \sin x$ is close to a third-degree algebraic polynomial on the interval under consideration.

We set in (4) $n = 3$, $x_0 = 5°$, $h = 7° - 5° = 2°$, and $q = (6° - 5°)/2° = 1/2$. The relevant computations have the form

$$f_0 = 0.087156,$$

$$q\Delta f_0 = 0.5 \cdot 0.034713 = 0.0073565,$$

$$q(q-1)\frac{\Delta^2 f_0}{2!} = \frac{1}{8} \cdot 0.000148 = 0.0000185,$$

$$q(q-1)(q-2)\frac{\Delta^3 f_0}{3!} = -\frac{1}{16} \cdot 0.000042 = -0.0000026,$$

$$l_3(6°) = 0.104528.$$

Here, the intermediate quantities are found with seven digits after the decimal point. The seventh digit is a spare one and in the final result it is rounded off.

The exact value of sin 6° rounded off to six decimal places is equal to 0.104528, that is, all digits of $l_3(6°)$ turn out to be correct.

The interpolation polynomial with points $x_0, x_{-1}, \ldots, x_{-n}$, where $x_{-k} = x_0 - kh$, has the form

$$l_n(x) = l_n(x_0 + qh) = f_0 + q\frac{\Delta f_{-1}}{1!} + q(q+1)\frac{\Delta^2 f_{-2}}{2!} + \ldots + q(q+1)\ldots(q+n-1)\frac{\Delta^n f_{-n}}{n!}. \quad (5)$$

It is called *Newton's interpolation polynomial for backward interpolation*. In it, the reference point q is situated at then rightmost interpolation point x_0, and the finite differences used are contained in the table from f_0 to the right upwards:

x_{-4}	f_{-4}			
		Δf_{-4}		
x_{-3}	f_{-3}		$\Delta^2 f_{-4}$	
		Δf_{-3}		$\Delta^3 f_{-4}$
x_{-2}	f_{-2}		$\Delta^2 f_{-3}$	
		Δf_{-2}		$\Delta^3 f_{-3}$
x_{-1}	f_{-1}		$\Delta^2 f_{-2}$	
		Δf_{-1}		
x_0	f_0			

The interpolation polynomial (5) may be conveniently used for interpolation at the end of the table and for extrapolation to the right of the point x_0, that is, for $q > 0$.

If for a given x the table of the values of the function f with step h has a sufficient number of interpolation points on either side of x, then, according to Remark 7.1, it is convenient to choose the interpolation points $x_0, x_1, \ldots, x_n$ so that the point x is found as close as possible to the middle of the minimum interval containing these points. In this case, the interpolation polynomial can be constructed in different ways. The most natural way to specify the interpolation polynomial in the form (1) where the nearest to x point is taken as x_0 (on one side of x) and then the nearest to x point is taken as x_1 (on the opposite side). The subsequent interpolation points are taken by turns on opposite sides of x and are situated as close to x as possible. With such a choice of the interpolation points, the subsequent terms in the expression (1) are usually decreasing if h is small and n is not large.

In the case under consideration, it is also possible to make use of the interpolation polynomials (4) and (5), as well as Lagrange's interpolation polynomial (7.4).

In conclusion, we should like to indicate that the remainder of the interpolation polynomial (4) has the expression (7.7), and the remainder of the interpolation polynomial (5) may be written in the form

$$R_n(x) = R_n(x_0 + qh) = h^{n+1} q(q+1)\ldots(q+n)\frac{f^{(n+1)}(\xi)}{(n+1)!},$$

where $f^{(n+1)}$ is the derivative with respect to x, and ξ is some point of the minimum interval containing the interpolation points $x_0, x_{-1}, \ldots, x_{-n}$ and the point x.

According to (8.4) and (7.7), if h is small and the function f is sufficiently smooth, then the current term in the expression (4) of Newton's interpolation polynomial is approximately equal to the error of interpolation by the polynomial made up of all the previous terms. This remark also refers to the interpolation polynomial (5) for backward interpolation.

Sec. 10. Numerical Differentiation

Below, we shall need the following lemma.

Lemma 1. *Let* $f \in C[a, b]$ and $\xi_i \in [a, b]$ be arbitrary points, $i = 1, 2, \ldots, n$. Then there is a point $\xi \in [a, b]$ such that*

$$\frac{f(\xi_1) + \ldots + f(\xi_n)}{n} = f(\xi).$$

* The class of functions continuous on the interval $[a, b]$ is denoted by $C[a, b]$; see Item 2 of Introduction.

Lemma 1 follows from the obvious inequalities

$$\min_{[a,\,b]} f(x) \leqslant \frac{f(\xi_1) + \ldots + f(\xi_n)}{n} \leqslant \max_{[a,\,b]} f(x)$$

and the theorem on the intermediate values of a continuous function.

Simplest Formulas of Numerical Differentiation. Let us assume that at some point x the function f has the derivative

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

whose exact computation is either impossible or too complicated. Then it is natural to set

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

There arises the following question: What is the error, that is, what is the difference between the left- and right-hand members of this approximate equality? To obtain the quantitative estimates of the error, the sole fact of the existence of $f'(x)$ is insufficient. Therefore, when analyzing the error of approximate formulas of numerical differentiation, we usually demand that the given function have some derivative of a higher order than the desired derivative.

Let us consider three most widely used formulas of numerical differentiation. Let $x_i = x_0 + ih$, $i = 0, \pm 1, \ldots$, where $h > 0$ is the step. Let us set $f_i = f(x_i), f_i' = f'(x_i)$, and so forth.

Suppose that $f \in C_2[x_0, x_1]$. Then there is a point ξ such that

$$f_0' = \frac{f_1 - f_0}{h} - \frac{h}{2} f''(\xi), \quad x_0 < \xi < x_1. \tag{1}$$

If $f \in C_3[x_{-1}, x_1]$, then, in addition,

$$f_0' = \frac{f_1 - f_{-1}}{2h} - \frac{h^2}{6} f'''(\xi), \quad x_{-1} < \xi < x_1. \tag{2}$$

Under the condition that $f \in C_4[x_{-1}, x_1]$, we have

$$f_0'' = \frac{f_{-1} - 2f_0 + f_1}{h^2} - \frac{h^2}{12} f^{(4)}(\xi), \quad x_{-1} < \xi < x_1. \tag{3}$$

The point ξ in each of the formulas (1)-(3) is unknown.

Let us prove the relationships (1) and (3). According to Taylor's formula, we have

$$f_1 = f_0 + hf_0' + \frac{h^2}{2} f''(\xi),$$

where ξ is some point of the interval (x_0, x_1), that is, (1) holds true.

Analogously, if $f \in C_4[x_{-1}, x_1]$, then

$$f_{\pm 1} = f_0 \pm hf_0' + \frac{h^2}{2} f_0'' \pm \frac{h^3}{6} f_0''' + \frac{h^4}{24} f^{(4)}(\xi_\pm), \tag{4}$$

where $\pm$ may be replaced either everywhere by $+$ or everywhere by $-$, $x_{-1} < \xi_- < \xi_+ < x_1$.

Substituting the values of f_{-1} and f_1 in the form (4) into the expression $(f_{-1} - 2f_0 + f_1)/h^2$ and taking into consideration that, by Lemma 1, $f^{(4)}(\xi_-) + f^{(4)}(\xi_+) = 2f(\xi)$, where ξ is some point of the interval $[\xi_-, \xi_+]$, we obtain

$$\frac{f_{-1} - 2f_0 + f_1}{h^2} = f_0'' + \frac{h^2}{12} f^{(4)}(\xi), \quad x_{-1} < \xi < x_1,$$

that is, we arrive at the relationship (3).

The formulas (1)-(3) are called the *formulas of numerical differentiation with remainders,* and the formulas

$$f_0' \approx \frac{f_1 - f_0}{h}, \tag{5}$$

$$f_0' \approx \frac{f_1 - f_{-1}}{2h}, \tag{6}$$

$$f_0'' \approx \frac{f_{-1} - 2f_0 + f_1}{h^2} \tag{7}$$

simply the *formulas of numerical differentiation.*

The quantities

$$\frac{f_1 - f_0}{h}, \quad \frac{f_1 - f_{-1}}{2h}, \quad \frac{f_{-1} - 2f_0 + f_1}{h^2}$$

are termed the *difference derivative, central difference derivative,* and the *second difference derivative,* respectively.

The errors, say, of the formulas (5) and (6) are estimated with the aid of the following inequalities which follow from the relationships (1) and (2):

$$\left| f_0' - \frac{f_1 - f_0}{h} \right| \leqslant \frac{h}{2} \max_{[x_0, x_1]} |f''(x)|,$$

$$\left| f_0' - \frac{f_1 - f_{-1}}{2h} \right| \leqslant \frac{h^2}{6} \max_{[x_{-1}, x_1]} |f'''(x)|.$$

The error of the formula (5) is said to have the *first order with respect to h* (or *to be of order h*), while the error of the formulas (6) and (7) is said to have the *second order with respect to h* (or *to be of order* h^2). The formula of numerical differentiation (5) is said to be of the *first order of accuracy* (with respect to h), while the formulas (6) and (7) are said to have the *second order of accuracy*.

Applying Lagrange's Interpolation Polynomial. There exist formulas of numerical differentiation of any order of accuracy for finding derivatives of any order. One of the general-purpose methods of constructing the formulas of numerical differentiation consists in the following: using the values of the function f at some points $x_0, x_1, \ldots, x_n$, we construct the interpolation polynomial $L_n(x)$ (4.5) and approximately set:

$$f^{(m)}(x) \approx L_n^{(m)}(x), \quad 0 \leqslant m \leqslant n. \tag{8}$$

In a number of cases, along with the approximate equality (8), we succeed in obtaining an exact equality containing a remainder expressed in terms of the derivative $f^{(n+1)}$. In addition to (1)-(3), let us give several widespread formulas for the first ($m = 1$) and second ($m = 2$) derivatives at the interpolation points spaced at a constant step $h > 0$.

$m = 1$ and $n = 2$ (three points):

$$f_0' = \frac{1}{2h}(-3f_0 + 4f_1 - f_2) + \frac{h^2}{3} f'''(\xi), \tag{9}$$

$$f_1' = \frac{1}{2h}(f_2 - f_0) - \frac{h^2}{6} f'''(\xi),$$

$$f_2' = \frac{1}{2h}(f_0 - 4f_1 + 3f_2) + \frac{h^2}{3} f'''(\xi). \tag{10}$$

$m = 2$ and $n = 2$ (three points):

$$f_0'' = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - hf'''(\xi), \tag{11}$$

$$f_1'' = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - \frac{h^2}{12}f^{(4)}(\xi),$$

$$f_2'' = \frac{1}{h^2}(f_0 - 2f_1 + f_2) + hf''(\xi).$$

$m = 1$ and $n = 3$ (four points):

$$f_0' = \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3) - \frac{h^3}{4}f^{(4)}(\xi), \tag{12}$$

$$f_1' = \frac{1}{6h}(-2f_0 - 3f_1 + 6f_2 - f_3) + \frac{h^3}{12}f^{(4)}(\xi),$$

$$f_2' = \frac{1}{6h}(f_0 - 6f_1 + 3f_2 + 2f_3) - \frac{h^3}{12}f^{(4)}(\xi),$$

$$f_3' = \frac{1}{6h}(-2f_0 + 9f_1 - 18f_2 + 11f_3) + \frac{h^3}{4}f^{(4)}(\xi). \tag{13}$$

$m = 2$ and $n = 3$ (four points):

$$f_0'' = \frac{1}{h^2}(2f_0 - 5f_1 + 4f_2 - f_3) + \frac{11}{12}h^2f^{(4)}(\xi),$$

$$f_1'' = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - \frac{h^2}{12}f^{(4)}(\xi),$$

$$f_2'' = \frac{1}{h^2}(f_1 - 2f_2 + f_3) - \frac{h^2}{12}f^{(4)}(\xi),$$

$$f_3'' = \frac{1}{h^2}(-f_0 + 4f_1 - 5f_2 + 2f_3) + \frac{11}{12}h^2f^{(4)}(\xi). \tag{14}$$

In the above formulas, ξ is an unknown point from the interval (x_0, x_n). The remainders of these formulas are found with the aid of Taylor's formula with the remainder in integral form. It is assumed here that on the interval $[x_0, x_n]$ the function f has a continuous derivative in terms of which the remainder is expressed.

We might add to the above list more formulas of numerical differentiation for increasing values of n and m. The following rules are then revealed: with an increase in n and an appropriate smoothness of the function f, the order of accuracy of the formulas is increased, and with an increase in m (that is, in the derivative number) the order of accuracy with respect to h decreases. The expressions of the derivatives at the points situated closer to the middle of the interval $[x_0, x_n]$ are simpler than those at its ends. For an even n the order of accuracy of the formula at the middle point for an even derivative is a unity higher than at the remaining points. Therefore it is recommended, as far as possible, to use the formulas of numerical differentiation with points arranged symmetrically about the point at which the derivative is determined.

The Choice of an Optimal Step. In formulas of numerical differentiation with constant step h the values of the function f are divided by h^m, where m is the order of the computed derivative. Therefore for a small h the inherent errors in the values of the function f strongly affect the result of numerical differentiation. Thus, there arises the problem of the choice of an optimal step h, since the error of the method itself tends to zero as $h \to 0$, and the inherent error increases.

Suppose that the absolute error $\Delta(f_i)$ of each value of the function f_i satisfies the inequality

$$\Delta(f_i) \leqslant \bar{\Delta}, \tag{15}$$

that is, $\bar{\Delta}$ is the limiting absolute error of the function values. Let us try to find the optimal h in the formulas (6) and (7). Let in some neighbourhood of the point x_0 the derivatives in terms of which the remainders of these formulas (see (2) and (3)) are expressed be continuous and satisfy the inequalities

$$|f'''(x)| \leqslant \bar{M}_3, \quad |f^{(4)}(x)| \leqslant \bar{M}_4, \tag{16}$$

where $\bar{M}_3$ and $\bar{M}_4$ are some numbers. Then the total error of the formulas (6) and (7) (neglecting the round-off errors), in accordance with (2), (3) and (15), (16), does not exceed the respective values:

$$\varepsilon_1 = \frac{\bar{\Delta} + \bar{\Delta}}{2h} + \frac{h^2}{6}\bar{M}_3, \tag{17}$$

$$\varepsilon_2 = \frac{\bar{\Delta} + 2\bar{\Delta} + \bar{\Delta}}{h^2} + \frac{h^2}{12}\bar{M}_4. \tag{18}$$

Minimizing these quantities with respect to h yields the following values of h:

$$h = h_1 = \left(\frac{3\bar{\Delta}}{\bar{M}_3}\right)^{1/3}, \quad h = h_2 = 2\left(\frac{3\bar{\Delta}}{\bar{M}_4}\right)^{1/4}, \tag{19}$$

and

$$\varepsilon_1 = \frac{3}{2}\left(\frac{\bar{M}_3\bar{\Delta}^2}{3}\right)^{1/3}, \quad \varepsilon_2 = 2\left(\frac{\bar{M}_4\bar{\Delta}}{3}\right)^{1/2}. \tag{20}$$

If for the value of h chosen for some of the formulas (6) or (7) the interval $[x_{-1}, x_1]$, where $x_{\pm 1} = x_0 \pm h$, does not fall outside the limits of the neighbourhood of the point x_0 at which the appropriate inequality (16) is fulfilled, then the found h is *optimal,* and the total error of numerical differentiation is estimated by the corresponding quantity (20). Otherwise, h is finally chosen so that the interval $[x_{-1}, x_1]$ does not fall outside the limits of the indicated neighbourhood of the point x_0, and the total error of numerical differentiation is estimated by (17) or (18).

Applications of Newton's Interpolation Polynomial. All the listed formulas of numerical differentiation are expressed in terms of the tabular values of the given function. If we differentiate Newton's interpolation polynomial (9.4), then we obtain the formula of numerical differentiation expressed in terms of finite differences of the function. Taking into consideration that, according to (7.1),

$$\frac{d}{dx} = \frac{1}{h}\frac{d}{dq}, \tag{21}$$

from (9.4) we find

$$f'(x) \approx \frac{1}{h}\frac{d}{dq} l_n(x_0 + qh)$$

$$= \frac{1}{h}\left(\Delta f_0 + (2q - 1)\frac{\Delta^2 f_0}{2!} + (3q^2 - 6q + 2)\frac{\Delta^3 f_0}{3!} + \ldots\right).$$

This formula is convenient to be used for interpolation at the beginning of the table of the values of the function f with step h. A similar formula of numerical differentiation can be obtained from Newton's interpolation polynomial (9.5) for backward interpolation.

General Error Estimate. The exact equalities of the forms (1)-(3) and (9)-(14) with remainder expressed in terms of $(n + 1)$th derivative of the function f turn out to be found easily for the formulas of numerical differentiation only in particular cases some of which were given above. But

it is possible to obtain the error estimate of the general formula of numerical differentiation (8) which is expressed in the form of an inequality in terms of the modulus maximum of the derivative $f^{(k+1)}$ for any m, k, and n such that $0 \leqslant m \leqslant k \leqslant n$. We shall confine ourselves to considering the case when the interpolation points are spaced at a constant step h and formulate the result in the form of a theorem.

Theorem 1. *Let $x_i = x_0 + ih$, $h > 0$, $i = 0, 1, \ldots, n$, $0 \leqslant k \leqslant n$, and $f \in C_{k+1}[x_0, x_n]$. Then there are constants a_{nkm} dependent only on n, k, and m and independent of the step h and the function f such that*

$$\max_{[x_0, x_n]} |f^{(m)}(x) - L_n^{(m)}(x)| \leqslant h^{k+1-m} a_{nkm} \max_{[x_0, x_n]} |f^{(k+1)}(x)|, \quad (22)$$

where $L_n(x)$ is Lagrange's interpolation polynomial (4.5) for the function f, $0 \leqslant m \leqslant k \leqslant n$.

Proof. The polynomial

$$Q_k(x) = f_0 + (x - x_0)\frac{f_0'}{1!} + \ldots + (x - x_0)^k \frac{f_0^{(k)}}{k!}$$

is Taylor's polynomial of degree k for the function f at the point x_0. Differentiating it m times, $0 \leqslant m \leqslant k$, we obtain the polynomial

$$Q_k^{(m)}(x) = f_0^{(m)} + (x - x_0)\frac{f_0^{(m+1)}}{1!} + \ldots + (x - x_0)^{k-m}\frac{f_0^{(k)}}{(k-m)!},$$

which is Taylor's polynomial of degree $k - m$ for the derivative $f^{(m)}$ at the point x_0.

Since $f^{(m)} \in C_{k+1-m}[x_0, x_n]$, according to Taylor's formula, we have for $x \in [x_0, x_n]$

$$f^{(m)}(x) = Q_k^{(m)}(x) + (x - x_0)^{k+1-m}\frac{f^{(k+1)}(\xi)}{(k+1-m)!}, \quad (23)$$

where $0 \leqslant m \leqslant k$ and $\xi \in [x_0, x]$ is dependent on x, m, and n.

If $x \in [x_0, x_n]$, then $|x - x_0| \leqslant nh$. Whence and from (23), for $0 \leqslant m \leqslant k$ there follows the estimate

$$\max_{[x_0, x_n]} |f^{(m)}(x) - Q_k^{(m)}(x)| \leqslant h^{k+1-m}\frac{n^{k+1-m}}{(k+1-m)!}\max_{[x_0, x_n]} |f^{(k+1)}(x)|. \quad (24)$$

We now set

$$f(x) = Q_k(x) + \rho(x). \tag{25}$$

Then, on the basis of (23), for $m = 0$ we shall have

$$|\rho_i| \leqslant h^{k+1} \frac{i^{k+1}}{(k+1)!} \max_{[x_0, x_n]} |f^{(k+1)}(x)|, \tag{26}$$

where $\rho_i = \rho(x_i)$, $i = 0, 1, \ldots, n$.

For the sake of convenience, let us denote the interpolation polynomial (4.5) just for the function f by $L_n(x; f)$. According to (25) and to Remarks 4.2 and 4.1,

$$L_n(x; f) = L_n(x; Q_k) + L_n(x; \rho) = Q_k(x) + L_n(x; \rho). \tag{27}$$

Introducing the dimensionless variable $q = (x - x_0)/h$, analogously to (7.4), we get

$$L_n(x; \rho) = \sum_{i=0}^{n} \bar{p}_{ni}(q)\, \rho_i, \tag{28}$$

where $\bar{p}_{ni}(q)$ are the Lagrangian coefficients (7.5). Taking into account (21), from (28) we find

$$L_n^{(m)}(x; \rho) = \frac{1}{h^m} \sum_{i=0}^{n} \rho_i \frac{d^m}{dq^m} \bar{p}_{ni}(q), \tag{29}$$

where the mth derivative on the left is taken with respect to x. Since the functions (7.5) are independent of h, the quantities

$$\max_{x \in [x_0, x_n]} \left| \frac{d^m}{dq^m} \bar{p}_{ni}(q) \right| = \max_{q \in [0, n]} \left| \frac{d^m}{dq^m} \bar{p}_{ni}(q) \right| = c_{nmi} \tag{30}$$

are dependent only on n, m, and i.

Returning to the original notation $L_n(x)$ of the interpolation polynomial for the function f and bearing in mind (27), we obtain

$$|f^{(m)}(x) - L_n^{(m)}(x)| = |f^{(m)}(x) - Q_k^{(m)}(x) - L_n^{(m)}(x; \rho)|$$

$$\leqslant |f^{(m)}(x) - Q_k^{(m)}(x)| + |L_n^{(m)}(x; \rho)|.$$

Hence, on the basis of (24), (29), (26), and (30) there follows the inequality (22) with the constant

$$a_{nkm} = \frac{n^{k+1-m}}{(k+1-m)!} + \frac{1}{(k+1)!} \sum_{i=1}^{n} i^{k+1} c_{nmi} \qquad (31)$$

independent of the step h and the function f.

Theorem 1 has been proved.

Remark 1. This theorem also yields additional information for $m = 0$, that is, simply for interpolation. Namely, if we construct an nth-degree interpolation polynomial and the function f has a continuous derivative only of order $k + 1$, $k < n$, on the interval $[x_0, x_n]$, then the error estimate (7.8) obtained earlier is unfit. But, according to Theorem 1, in this case the interpolation error on the interval $[x_0, x_n]$ has the $(k + 1)$th order with respect to h.

Remark 2. In practice, it is inappropriate to use the estimate (22) with the constant (31) for numerical differentiation, since this constant is too excessive. The estimate (22) is useful, since it establishes the rate of the error decrease relative to h on the entire interval $[x_0, x_n]$ for fixed parameters n, k, and m, $0 \leqslant m \leqslant k \leqslant n$, the step h being one of the most flexible parameters at computer's disposal.

Sec. 11. Splines

Let the interval $[a, b]$ be divided into N equal subintervals $[x_i, x_{i+1}]$, where $x_i = a + ih$, $i = 0, 1, \ldots, N - 1$, $x_N = b$, and $h = (b - a)/N$.

The *spline* is defined as a function which, together with its several derivatives, is continuous throughout the given interval $[a, b]$ and on each separate subinterval $[x_i, x_{i+1}]$ it is some algebraic polynomial.

The maximum (over all the subintervals) degree of polynomials is called the *degree* of a spline, and the difference between the degree of a spline and the order of the highest derivative continuous on $[a, b]$ is called the *defect* of a spline.

For instance, a continuous piecewise linear function (a broken line) is a first-degree spline with defect equal to unity, since only the function itself (or zero derivative) is continuous, while the first derivative is already discontinuous.

In practice, the most common are third-degree splines having a continuous, at least, first derivative on $[a, b]$. These are called *cubic splines* and are denoted by $S_3(x)$. The quantity $m_i = S_3'(x_i)$ is termed the *inclination* of the spline at the point (knot) x_i.

It is not difficult to make sure that the cubic spline $S_3(x)$ which attains the respective values f_i and f_{i+1} at the knots x_i and x_{i+1} has the following expression on the subinterval $[x_i, x_{i+1}]$:

$$S_3(x) = \frac{(x_{i+1} - x)^2(2(x - x_i) + h)}{h^3} f_i + \frac{(x - x_i)^2(2(x_{i+1} - x) + h)}{h^3} f_{i+1} + \frac{(x_{i+1} - x)^2(x - x_i)}{h^2} m_i + \frac{(x - x_i)^2(x - x_{i+1})}{h^2} m_{i+1}. \quad (1)$$

Indeed, it is readily seen that $S_3(x_i) = f_i$ and $S_3(x_{i+1}) = f_{i+1}$. Besides, simple computations show that $S_3'(x_i) = m_i$ and $S_3'(x_{i+1}) = m_{i+1}$.

It is possible to prove that any third-degree algebraic polynomial, which attains the values respectively equal to f_i and f_{i+1} at the knots x_i and x_{i+1} and has a derivative respectively equal to m_i and m_{i+1} at these points, coincides identically with the polynomial (1).

Thus, to define the cubic spline $S_3(x)$ over the whole interval $[a, b]$, we have to specify its values f_i at $N + 1$ knots x_i and its inclinations m_i, $i = 0, 1, \ldots, N$.

Splines are used for various purposes. A cubic spline which attains the same values f_i at the knots x_i as a function f is called the *interpolation spline*. It serves for approximating the function f on the interval $[a, b]$ together with several derivatives.

Methods of Specifying the Inclinations of an Interpolation Cubic Spline.

Method 1 (*simplified*). We set

$$m_i = \frac{f_{i+1} - f_{i-1}}{2h}, \quad i = 1, 2, \ldots, N - 1, \quad (2)$$

$$m_0 = \frac{4f_1 - f_2 - 3f_0}{2h}, \quad m_N = \frac{3f_N + f_{N-2} - 4f_{N-1}}{2h}. \quad (3)$$

Note that, according to (10.2), (10.9), and (10.10), the formulas (2) and (3) are formulas of numerical differentiation of the second order of accuracy with respect to the step $h = (b - a)/N$.

Method 2. If the values f_i' of the derivative f' at the knots x_i are known, then we set $m_i = f_i'$, $i = 0, 1, \ldots, N$.

Methods 1 and 2 are called *local*, since with their aid the spline is constructed separately on each subinterval $[x_i, x_{i+1}]$ (directly by the formula (1)). In doing so, nevertheless, the continuity of the derivative $S_3'(x)$ at the knots x_i is observed. But the continuity of the second derivative $S_3''(x)$ at the knots of the spline constructed by local method 1 or 2 is not guaranteed. Therefore the defect of such a spline is usually equal to two.

Method 3 (*global*). We denote by $S_3''(x_i + 0)$ the value of $S_3''(x)$ at the knot x_i from the right, i.e. the value found directly via the expression (1), and by $S_3''(x_i - 0)$ the value of $S_3''(x)$ at the knot x_i from the left, i.e. the value found via the appropriate expression of $S_3(x)$ on the subinterval $[x_{i-1}, x_i]$ which is obtained from (1) by substituting $i - 1$ for i. We have

$$S_3''(x_i + 0) = -\frac{4m_i}{h} - \frac{2m_{i+1}}{h} + 6\frac{f_{i+1} - f_i}{h^2},$$

$$S_3''(x_i - 0) = \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - 6\frac{f_i - f_{i-1}}{h^2}.$$

Then the continuity of $S''(x)$ is required at the knots:

$$S_3''(x_i - 0) = S_3''(x_i + 0), \quad i = 1, 2, \ldots, N - 1.$$

We arrive at the following system of linear algebraic equations with respect to inclinations:

$$m_{i-1} + 4m_i + m_{i+1} = \frac{3(f_{i+1} - f_{i-1})}{h}, \quad i = 1, 2, \ldots, N - 1. \quad (4)$$

Since there are $N + 1$ unknowns, it is necessary to specify two more conditions which are called the *boundary conditions* (they are usually connected with the "extreme" values m_0 and m_N). Let us give three variants of boundary conditions.

(a) With $f_0' = f'(a)$ and $f_N' = f'(b)$ known, we specify

$$m_0 = f_0', \quad m_N = f_N'. \quad (5)$$

(b) The derivatives f_0' and f_N' are approximated by the formulas of numerical differentiation of the third order of accuracy from (10.12) and (10.13) by rejecting the remainder. We set

$$m_0 = \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3),$$
$$m_N = \frac{1}{6h}(11f_N - 18f_{N-1} + 9f_{N-2} - 2f_{N-3}). \quad (6)$$

(c) In some cases the values of f'' at the end points of the interval $[a, b]$ are known, that is, the quantities $f_0'' = f''(a)$ and $f_N'' = f''(b)$. Then the requirements $S_3''(a) = f_0''$ and $S_3''(b) = f_N''$ lead to the boundary conditions

$$m_0 = -\frac{m_1}{2} + \frac{3}{2}\frac{f_1 - f_0}{h} - \frac{h}{4} f_0'',$$
$$m_N = -\frac{m_{N-1}}{2} + \frac{3}{2}\frac{f_N - f_{N-1}}{h} + \frac{h}{4} f_N''. \tag{7}$$

The boundary conditions (5)-(7) may be combined, that is, we may choose them independently at the left and right extreme knots.

For all considered boundary conditions, the system (4) has a unique solution which can be found by applying the sweep and iteration methods (this will be proved in Secs. 21 and 22).

Solving the system (4) for the chosen boundary conditions, we find the inclination m_i, $i = 0, 1, \ldots, N$, at all the knots. Then, making use of the formula (1), we specify the spline on each subinterval $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, N-1$. The spline $S_3(x)$ constructed by this global method has a defect not exceeding unity, since this spline has a continuous second derivative $S_3''(x)$ on the interval $[a, b]$.

The Error of Approximation by a Spline. The cubic interpolation spline yields a good approximation of a smooth function together with several derivatives, which is testified by the following theorem (the latter is proved analogously to Theorem 10.1).

Theorem 1. *If $f \in C_{k+1}[a, b]$, $0 \leqslant k \leqslant 3$, then the interpolation spline $S_3(x)$, with the inclinations given by Method 2 or 3, satisfies the inequality*

$$\max_{[x_i, x_{i+1}]} |f^{(m)}(x) - S_3^{(m)}(x)| \leqslant ch^{k+1-m} \max_{[a,b]} |f^{(k+1)}(x)|, \tag{8}$$

where $i = 0, 1, \ldots, N-1$, $m = 0, 1, \ldots, k$, and c is a constant independent of h, i, and f.

Thus, if $f \in C_4[a, b]$, i.e. $k = 3$, and the inclinations of the spline $S_3(x)$ are found by the global method, then on $[a, b]$ the maximum-modulus deviations of $S_3(x)$ from $f(x)$, $S_3'(x)$ from $f'(x)$, and $S_3''(x)$ from $f''(x)$ are equal to $O(h^4)$, $O(h^3)$, and $O(h^2)$, respectively. When specifying the inclinations by Method 2, the jumps of $S_3''(x)$ occurring at the knots do not exceed the doubled right-hand member of the inequality (8), that is, for $k = 3$ they will be equal to $O(h^2)$. The jumps of the second derivative are difficult to be observed on the graph.

Remark 1. For the cubic spline $S_3(x)$ whose inclinations are given by the simplified method 1 Theorem 1 holds true provided that $0 \leqslant k \leqslant 2$.

Remark 2. If the function f is continuously differentiable throughout the real axis $k + 1$ times, $0 \leqslant k \leqslant 3$, and has a period equal to $b - a$, we have to set $m_0 = m_N$ and to the system (4) we join the equation

$$m_{N-1} + 4m_0 + m_1 = \frac{3(f_1 - f_{N-1})}{h}$$

which corresponds to the value $i = 0$. We find the inclinations of the spline $S_3(x) \in C_2[a, b]$ from the extended system which has a unique solution. This spline is continued from the interval $[a, b]$ with period $b - a$ onto the whole x-axis, the continuity of the second derivative being retained. In this case, the inequalities (8) are valid.

Splines are a more convenient means of approximation of functions on big intervals (for large N) than, say, interpolation polynomials. An approximation of a function on a big interval by one polynomial may require a considerable increase in its degree in order to achieve the assigned accuracy, which is unacceptable in practice. The subdivision of the given interval $[a, b]$ into several parts with an interpolation polynomial constructed independently on each subinterval is inconvenient, since the first derivative of two neighbouring interpolation polynomials will have a discontinuity. It may even happen that the values themselves of two adjacent interpolation polynomials will not coincide at the joint if the joint point is not their common knot. The cubic spline $S_3(x)$ whose inclinations are found by the global method is twice continuously differentiable on the entire interval $[a, b]$, that is, it has a continuous curvature. The accuracy of the approximation of the function f by the spline $S_3(x)$ is controlled by the choice of N, that is, by the step $h = (b - a)/N$.

Sec. 12. Uniform Approximation of Functions

In the linear space of continuous functions $C[a, b]$, let us introduce the norm*

$$\|f\|_{C[a,b]} = \max_{[a,b]} |f(x)|. \tag{1}$$

The norm of the difference of two functions $f, g \in C[a, b]$, i.e. the quantity

$$\|f - g\|_{C[a,b]} = \max_{[a,b]} |f(x) - g(x)|,$$

* The concepts of the norm and normed space are given in Item 7 of Introduction.

is equal to the maximum deviation of these functions from each other on the interval $[a, b]$.

Let a function f belong to $C[a, b]$. It was proved that among all algebraic polynomials of degree n there exists the only polynomial $P_n^f(x)$ having the property that

$$\|f - P_n^f\|_{C[a, b]} \leqslant \|f - P_n\|_{C[a, b]},$$

where $P_n(x)$ is an arbitrary polynomial of degree n.

The polynomial $P_n^f(x)$ is called the *polynomial (of degree n) of best uniform approximation* of the function f. The number

$$E_n(f) = \|f - P_n^f\|_{C[a, b]} = \inf \|f - P_n\|_{C[a, b]},$$

where the greatest lower bound is taken over all polynomials of degree n, is called the *best uniform approximation* of the function f by polynomials of degree n. Obviously, $E_n(f) \leqslant E_{n-1}(f)$ for all natural n.

It is possible to prove that whatever the function $f \in C[a, b]$, the sequence $\{P_n^f(x)\}$ of its polynomials of best uniform approximation converges uniformly to f on $[a, b]$, that is, $E_n(f) \to 0$ as $n \to \infty$.

Example. Let $f(x) = x^{n+1}$. It is required to find the nth-degree polynomial of best uniform approximation for the given function on the interval $[-1, 1]$.

Solution. Consider the polynomial

$$T_{n+1}(x) = 2^{-n} T_{n+1}(x), \tag{2}$$

where $T_{n+1}(x)$ is a Chebyshev polynomial of degree $n + 1$, least deviating from zero (see Sec. 6). According to Properties 1 and 2 of Chebyshev polynomials, the polynomial (2) can be written in the following form:

$$\bar{T}_{n+1}(x) = x^{n+1} - \alpha x^{n-1} - \beta x^{n-3} - \ldots, \tag{3}$$

where $\alpha, \beta, \ldots$ are quite definite numerical coefficients.

By Property 5 of Chebyshev polynomials, we have

$$\|\bar{T}_{n+1}\|_{C[-1, 1]} \leqslant \|\bar{P}_{n+1}\|_{C[-1, 1]}, \tag{4}$$

where $\bar{P}_{n+1}(x)$ is an arbitrary polynomial of degree $n + 1$ with leading coefficient equal to unity.

From (3) and (4) it follows that the polynomial of best uniform approximation $P_n^f(x)$ for the function $f(x) = x^{n+1}$ on the interval $[-1, 1]$ has the form

$$P_n^f(x) = \alpha x^{n-1} + \beta x^{n-3} + \ldots. \tag{5}$$

Indeed, $\bar{T}_{n+1}(x) = x^{n+1} - P_n^f(x) = f(x) - P_n^f(x)$ and $\bar{P}_{n+1}(x) = x^{n+1} - P_n(x) = f(x) - P_n(x)$, where $P_n(x)$ is an arbitrary polynomial of degree n and, according to (4), the required inequality

$$\|f - P_n^f\|_{C[-1, 1]} \leqslant \|f - P_n\|_{C[-1, 1]}$$

is fulfilled. And, on the basis of (2) and Property 4 of Chebyshev polynomials, we obtain

$$E_n(x^{n+1}) = E_n(f) = \|f - P_n^f\|_{C[-1, 1]} = \|\bar{T}_{n+1}\|_{C[-1, 1]}$$
$$= \|2^{-n} T_{n+1}\|_{C[-1, 1]} = 2^{-n} \|T_{n+1}\|_{C[-1, 1]} = 2^{-n}.$$

Criterion for a Polynomial of Best Uniform Approximation. Let us now pay attention to one important property of the polynomial of best uniform approximation $P_n^f(x)$ in the considered example. By virtue of Property 4 of Chebyshev polynomials, the difference $f(x) - P_n^f(x)$ which coincides with the expression (2) for $\bar{T}_{n+1}(x)$ attains the values $(-1)^m \|f - P_n^f\|_{C[-1, 1]}$ at $n + 2$ points $x_m = \cos(m\pi/(n + 1))$, $m = 0, 1, \ldots, n + 1$, of the interval $[-1, 1]$. In other words, the interval $[-1, 1]$ contains $n + 2$ points at which the difference $f(x) - P_n^f(x)$ reaches the maximum-modulus value with alternating signs.

The indicated property turns out to be general and serves as a criterion for the polynomial of best uniform approximation to any continuous function. The following theorem (given without proof) testifies to this fact.

Theorem 1 (Chebyshev's Theorem). *For a polynomial $P_n(x)$ to be a polynomial of best uniform approximation to a continuous function $f \in C[a, b]$, it is necessary and sufficient that there exists on $[a, b]$ at least $n + 2$ points $x_0 < \ldots < x_{n+1}$ such that*

$$f(x_i) - P_n(x_i) = \sigma(-1)^i \|f - P_n\|_{C[a, b]},$$

where $i = 0, 1, \ldots, n + 1$ and $\sigma = 1$ or -1 simultaneously for all i.

The points $x_0, \ldots, x_{n+1}$ which satisfy the conditions of the theorem are called the *points of Chebyshev's alternance* (or *critical points*).

Example. The function $f \in C[a, b]$ is approximated by a polynomial of zero degree. Let

$$\max_{[a, b]} f(x) = f(x_1) = M, \quad \max_{[a, b]} f(x) = f(x_2) = m.$$

The polynomial $P_0(x) = (M + m)/2$ is a polynomial of best uniform approximation (Fig. 3), x_1 and x_2 are critical points, and $E_0(f) = (M - m)/2$.

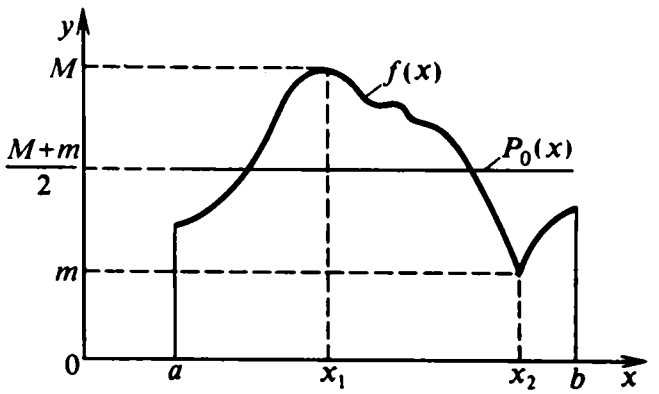


Fig. 3

Example. A convex function f which is differentiable on $[a, b]$ is approximated by the polynomial $P_1(x) = a_0 + a_1x$. Since the function f is convex, the difference $f(x) - (a_0 + a_1x)$ may have only one point of extremum on the interval (a, b). Therefore the points a and b are points of Chebyshev's alternance. Let $c \in (a, b)$ be a third point of the alternance. According to Chebyshev's theorem, the following conditions must be fulfilled:

$$f(a) - (a_0 + a_1a) = R, \quad f(c) - (a_0 + a_1c) = -R,$$
$$f(b) - (a_0 + a_1b) = R,$$

where $R = \sigma \|f - P_1^f\|_{C[a, b]}$ is unknown. From the first and third equations we get

$$a_1 = \frac{f(b) - f(a)}{b - a}.$$

The point c is the point of extremum of the difference $f(x) - (a_0 + a_1x)$ and is determined from the equation $f'(c) - a_1 = 0$. Further, from the first and second equations we determine a_0 and $E_1(f) = |R|$.

Geometrically, the solution of this problem is reduced to the following (Fig. 4). First we draw a chord through the points $(a, f(a))$ and $(b, f(b))$ with slope a_1. Then, with the same slope, we draw a tangent to the curve $y = f(x)$. Finally, we construct a secant in the middle between the chord and the tangent which yields the best uniform approximation of the curve $y = f(x)$ on the interval $[a, b]$ among all straight lines.

This geometrical method is obviously applied to the case of an arbitrary continuous function f whose graph is situated on one side of the segment joining the points $(a, f(a))$ and $(b, f(b))$.

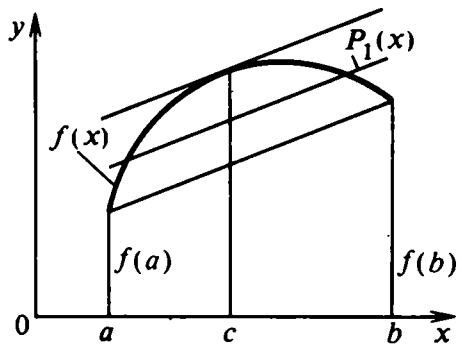


Fig. 4

In the general case, the construction of an nth-degree polynomial of best uniform approximation involves great difficulties. Therefore we often confine ourselves to finding a polynomial close to the best one or, in general, some polynomial $P_n(x)$ yielding a uniform approximation of the given function f with a required accuracy $\varepsilon > 0$, that is, satisfying the condition $\|f - P_n\|_{C[a, b]} \leqslant \varepsilon$.

Two Methods of Finding Polynomials Close to the Best Ones.

1. If $f \in C_{n+1}[a, b]$ and the derivative $f^{(n+1)}$ varies but little on $[a, b]$, then the interpolation polynomial $L_n(x)$ (4.5) with Chebyshev's points (6.11) is close to the best one. The value of deviation $\|f - L_n\|_{C[a, b]}$ is estimated by the inequality (6.12). In particular, if $f^{(n+1)}(x) = \text{const}$, then $L_n(x) \equiv P_n^f(x)$, and if $f(x) = x^{n+1}$ and $[a, b] = [-1, 1]$, then $L_n(x)$ coincides with the polynomial (5).

2. For a function expandable into a Taylor's series we may apply the method to be illustrated by an example. Let it be required to approximate uniformly the function $f(x) = \arctan x$ on the interval $[-\alpha, \alpha]$, where $\alpha = \tan(\pi/8)$, with an accuracy of $0.5 \cdot 10^{-5}$. We first take the truncated Taylor series which approximates the given function with somewhat higher accuracy than the assigned one. For instance, we set

$$f(x) \approx Q_{11}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11}.$$

Here,

$$\|f - Q_{11}\|_{C[-\alpha, \alpha]} < 0.82 \cdot 10^{-6}. \tag{6}$$

Via the substitution $x = \alpha t$, we pass to the interval $[-1, 1]$:

$$S_{11}(t) = Q_{11}(\alpha t)$$
$$= 0.4142136t - 0.0236893t^3 + 0.0024387t^5$$
$$- 0.0002989t^7 + 0.0000399t^9 - 0.0000056t^{11}.$$

Adding to the given polynomial the polynomial $0.0000056 \times \times 2^{-10} T_{11}(t)$, where $T_{11}(t)$ is the Chebyshev polynomial:

$$T_{11}(t) = 1024t^{11} - 2816t^9 + 2816t^7 - 1232t^5 + 220t^3 - 11t,$$

we obtain the polynomial of a lower degree

$$S_9(t) = S_{11}(t) + 0.0000056 \cdot 2^{-10} T_{11}(t)$$
$$= 0.4142135t - 0.0236881t^3 + 0.0024319t^5$$
$$- 0.0002835t^7 + 0.0000245t^9.$$

We have

$$\|S_{11} - S_9\|_{C[-1,1]} = \|0.0000056 \cdot 2^{-10} T_{11}\|_{C[-1,1]}$$
$$= \frac{0.0000056}{1024} \|T_{11}\|_{C[-1,1]} = \frac{0.0000056}{1024} < 0.01 \cdot 10^{-6}.$$

Thus, on the interval $[-1, 1]$ the polynomial $S_9(t)$ deviates from the polynomial $S_{11}(t)$ by the quantity which is considerably less than the maximum-modulus value of the leading term of the polynomial $S_{11}(t)$ for $t = \pm 1$. Therefore this method of reducing the degree of a polynomial is preferable as compared with mere rejection of the leading term.

Proceeding in a similar way two more times, we find

$$S_7(t) = S_9(t) - 0.0000245 \cdot 2^{-8} T_9(t)$$
$$= 0.4142127t - 0.0236766t^3$$
$$+ 0.0023906t^5 - 0.0002284t^7,$$
$$S_5(t) = S_7(t) + 0.0002284 \cdot 2^{-6} T_7(t)$$
$$= 0.4141877t - 0.0234768t^3 + 0.0019910t^5,$$

where $T_9(t)$ and $T_7(t)$ are the Chebyshev polynomials:

$$T_9(t) = 256t^9 - 576t^7 + 432t^5 - 120t^3 + 9t,$$
$$T_7(t) = 64t^7 - 112t^5 + 56t^3 - 7t.$$

And

$$\|S_9 - S_7\|_{C[-1,1]} = 0.0000245 \cdot 2^{-8} < 0.1 \cdot 10^{-6},$$
$$\|S_7 - S_5\|_{C[-1,1]} = 0.0002284 \cdot 2^{-6} < 0.36 \cdot 10^{-5}.$$

Returning to the original variable x, we obtain

$$P_5(x) = S_5\left(\frac{x}{\alpha}\right) = 0.9999374x - 0.3303433x^3 + 0.1632823x^5, \quad (7)$$

where

$$\|f - P_5\|_{C[-\alpha,\alpha]} \leqslant \|f - Q_{11}\|_{C[-\alpha,\alpha]} + \|S_{11} - S_9\|_{C[-1,1]} + \|S_9 - S_7\|_{C[-1,1]} + \|S_7 - S_5\|_{C[-1,1]} < 0.82 \cdot 10^{-6} + 0.01 \cdot 10^{-6} + 0.1 \cdot 10^{-6} + 0.36 \cdot 10^{-5} < 0.46 \cdot 10^{-5}.$$

Thus, we have found the polynomial (7) of fifth degree which approximates uniformly the function arc tan x on the interval $[-\tan(\pi/8), \tan(\pi/8)]$ with the assigned accuracy $0.5 \cdot 10^{-5}$. However, Taylor's polynomial of seventh degree does not ensure this accuracy.

The following simple theorem may turn out to be useful, its proof is left to the reader.

Teorem 2. *Let* $f \in C[-1, 1]$,

$$P_{n+1}(x) = a_0 + a_1x + \ldots + a_{n+1}x^{n+1}. \quad (8)$$

Then

$$E_n(f) \geqslant |a_{n+1}|\, 2^{-n} - \|f - P_{n+1}\|_{C[-1,1]}. \quad (9)$$

From this theorem we arrive at the following conclusion. Suppose that we have found a polynomial (8) which satisfies us, that is, it approximates uniformly the function f on the interval $[-1, 1]$ with accuracy ε, namely, $\|f - P_{n+1}\|_{C[-1,1]} \leqslant \varepsilon$. Then if the inequality

$$|a_{n+1}|2^{-n} - \|f - P_{n+1}\|_{C[-1,1]} > \varepsilon \quad (10)$$

is fulfilled, one should not try to seek for a polynomial of a lower degree which would ensure a uniform approximation of the function f with accuracy ε—there is none, since according to (9) and (10), $E_n(f) > \varepsilon$.

Let us apply Theorem 2 in the above-considered example. We have $\varepsilon = 0.5 \cdot 10^{-5}$, $n = 4$, and $\|f - P_5\|_{C[-\alpha,\alpha]} < 0.46 \cdot 10^{-5}$. When passing to the interval $[-1, 1]$, the leading coefficient of the polynomial (7) coincides with the leading coefficient of the polynomial S_5 equal to 0.0019910.

Since $0.0019910 \cdot 2^{-4} - 0.46 \cdot 10^{-5} > 0.5 \cdot 10^{-5}$, the inequality (10) is fulfilled. Consequently, even a fourth-degree polynomial of best uniform approximation will not be able to ensure the specified accuracy of approximation of the function arc tan x on the interval $[-\tan(\pi/8), \tan(\pi/8)]$. Thus, the polynomial (7) satisfying the given requirement has the minimum degree.

Sec. 13. The Method of Least Squares

Preliminaries. Consider the linear space F of real functions $f, g, \ldots$ defined on a set D of points x of the real axis (see Item 5 of Introduction). The zero element in F will be denoted by θ.

A *scalar product* is said to be introduced in the linear space F if to each pair of elements $f, g \in F$ there corresponds a real number denoted by (f, g) and called the scalar product of the elements f and g which satisfies the following *axioms of scalar product*:

(a) $(f, g) = (g, f)$;

(b) $(f, f) \geqslant 0$, *where* $(f, f) = 0$ *if and only if* $f = \theta$, *that is,* f *is a zero element in* F;

(c) $(\alpha f, g) = \alpha(f, g)$, *where* α *is any real number*;

(d) $(f_1 + f_2, g) = (f_1, g) + (f_2, g)$.

The linear space F with scalar product (f, g) introduced in it is called the *Euclidean space* and is denoted by E.

For definiteness, let us indicate that we shall be interested only in two concrete Euclidean spaces, namely, the space $E = E_C$ of functions which are continuous on the interval $[a, b]$ $(D = [a, b])$ with the scalar product

$$(f, g) = \frac{1}{b - a} \int_a^b f(x) g(x) dx \tag{1}$$

and the linear space $E = E_{n+1}$ of functions defined on a finite (discrete) set of points $x_0, x_1, \ldots, x_n$ of some interval $[a, b]$ $(D = \{x_i\}_{i=0}^n \subset [a, b])$ with the scalar product

$$(f, g) = \frac{1}{n + 1} \sum_{i=0}^{n} f(x_i) g(x_i). \tag{2}$$

The function $f \in E_{n+1}$ may be regarded as an $(n + 1)$-dimensional vector $f = (f_0, \ldots, f_n)$, where $f_i = f(x_i)$. The scalar product (2) differs from the scalar product in the $(n + 1)$-dimensional Euclidean vector space studied in the course of linear algebra only by the presence of an addi-

tional multiplier $1/(n+1)$. Therefore, obviously, all axioms of scalar product verified in linear algebra are fulfilled for it*.

Let us make sure that (1) is a scalar product in $C[a, b]$. Let us dwell on Axiom (b): for any function $f \in C[a, b]$

$$(f, f) = \frac{1}{b-a} \int_a^b f^2(x)\, dx \geqslant 0.$$

If $f = \theta$, i.e. $f(x) \equiv 0$, then, obviously, $(f, f) = 0$. Let us now assume, conversely, that $(f, f) = 0$, that is,

$$\int_a^b f^2(x)\, dx = 0. \tag{3}$$

Suppose that in this case $f(x) \not\equiv 0$ on $[a, b]$. Then by virtue of the continuity of the function f^2, there is an interval $[a', b'] \subset [a, b]$ on which $f^2(x) > 0$, and, consequently, $m = \min\limits_{[a', b']} f^2(x) > 0$, since a continuous function reaches its minimum on an interval. Hence,

$$\int_a^b f^2(x)\, dx \geqslant \int_{a'}^{b'} f^2(x)\, dx \geqslant \int_{a'}^{b'} m\, dx = m(b' - a') > 0.$$

But this contradicts (3). Therefore $f = \theta$.

Thus, Axiom (b) has been fulfilled. The verification of Axioms (c) and (d) for the scalar product (1) is readily carried out in terms of the known linear properties of the definite integral, and the fulfilment of Axiom (a) is obvious.

Note that *every Euclidean space is simultaneously a linear normed space* (see Item 7 of Introduction) *with the norm*

$$\|f\| = (f, f)^{1/2} \tag{4}$$

and, consequently (see Item 8 of Introduction), *a metric space with the distance*

$$\rho(f, g) = \|f - g\| = (f - g, f - g)^{1/2}. \tag{5}$$

For the norm introduced via a scalar product by the method (4), all the three axioms of norm formulated in Item 7 of Introduction are fulfilled.

---

* See *Fundamentals of Linear Algebra and Analytical Geometry* by Ya. S. Bugrov and S. M. Nikolsky, Sec. 6 (in English—Mir Publishers, Moscow, 1982).

Axiom (1) follows directly from Axiom (b) of scalar product. Let us verify Axiom (2).

Let α be any real number. We have $\|\alpha f\|^2 = (\alpha f, \alpha f) = \alpha(f, \alpha f) = \alpha(\alpha f, f) = \alpha^2(f, f) = \alpha^2\|f\|^2$. Hence, $\|\alpha f\| = |\alpha|\,\|f\|$, since $\|\alpha f\| \geqslant 0$. Axiom (2) has been fulfilled. Axiom (3), that is, the triangle inequality

$$\|f + g\| \leqslant \|f\| + \|g\|,$$

is established for the norm (4) in an arbitrary Euclidean space just in the same manner as Minkowski's inequality in an n-dimensional Euclidean vector space.

Let in the Euclidean space E there be given a system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Definition. The determinant

$$\begin{vmatrix} (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \ldots & (\varphi_m, \varphi_0) \\ (\varphi_0, \varphi_1) & (\varphi_1, \varphi_1) & \ldots & (\varphi_m, \varphi_1) \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots\ldots & \ldots\ldots\ldots \\ (\varphi_0, \varphi_m) & (\varphi_1, \varphi_m) & \ldots & (\varphi_m, \varphi_m) \end{vmatrix} \tag{6}$$

made up of scalar products is called the *Gram determinant* (or *Gramian*) of the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Lemma 1. *The Grum determinant* (6) *is equal to zero if and only if the system of functions* $\varphi_0, \varphi_1, \ldots, \varphi_m$ *is linearly dependent*.*

Proof. Suppose that the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is linearly dependent. Then there exist numbers $\alpha_0, \alpha_1, \ldots, \alpha_m$, not all zero, such that the equality

$$\alpha_0\varphi_0 + \alpha_1\varphi_1 + \ldots + \alpha_m\varphi_m = \theta \tag{7}$$

is fulfilled.

Multiplying this equality in turn by φ_0, φ_1, and so on and taking into account that** $(\theta, \varphi_j) = 0\ \forall\, j$, we obtain the relationships

$$\begin{aligned} &\alpha_0(\varphi_0, \varphi_0) + \alpha_1(\varphi_1, \varphi_0) + \ldots + \alpha_m(\varphi_m, \varphi_0) = 0, \\ &\alpha_0(\varphi_0, \varphi_1) + \alpha_1(\varphi_1, \varphi_1) + \ldots + \alpha_m(\varphi_m, \varphi_1) = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\alpha_0(\varphi_0, \varphi_m) + \alpha_1(\varphi_1, \varphi_m) + \ldots + \alpha_m(\varphi_m, \varphi_m) = 0, \end{aligned} \tag{8}$$

---

* For the definition of a linearly dependent and a linearly independent system of functions see Item 6 of Introduction.

** Since $\theta = 0 \cdot f\ \forall f \in E$, according to Axiom (c) of scalar product, we have $(\theta, f) = (0 \cdot f, f) = 0 \cdot (f, f) = 0\ \forall f \in E$.

which may be treated as a homogeneous system of linear algebraic equations having a nonzero solution $\alpha_0, \alpha_1, \ldots, \alpha_m$. Consequently, the determinant of the system (8) coinciding with the Gram determinant (6) is equal to zero.

Now, suppose conversely that the Gram determinant (6) is equal to zero. Let us prove that the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is linearly dependent. Since the determinant of the system (8) is equal to zero, the system has a certain nonzero solution $\alpha_0 = \bar{\alpha}_0, \alpha_1 = \bar{\alpha}_1, \ldots, \alpha_m = \bar{\alpha}_m$. Substituting the values $\alpha_j = \bar{\alpha}_j$, $j = 0, 1, \ldots, m$, into each equation of the system (8) and making use of Properties (Axioms) (a), (c), and (d) of scalar product, we obtain the following equalities:

$$\begin{aligned}
&(\varphi_0, \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m) = 0,\\
&(\varphi_1, \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(\varphi_m, \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m) = 0.
\end{aligned} \tag{9}$$

Multiplying these equalities by $\bar{\alpha}_0, \bar{\alpha}_1, \ldots, \bar{\alpha}_m$, respectively, and adding together the obtained results, we find

$$\|\bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m\|^2 = 0.$$

Consequently, by Axiom (1) of norm, we have

$$\bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m = \theta. \tag{10}$$

Since not all the numbers $\bar{\alpha}_j$ are zero, the equality (10) implies that the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is linearly dependent.

Lemma 1 has been proved.

Definition. A system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is said to be *orthogonal* if

$$(\varphi_j, \varphi_k) = 0, \quad j \neq k \quad (\varphi_j, \varphi_j) > 0, \tag{11}$$

where $0 \leqslant j$ and $k \leqslant m$.

Obviously, *if the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is orthogonal, then it is linearly independent.* Indeed, let us assume that the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is orthogonal, and for it the equality (7) is fulfilled, where $\alpha_0, \alpha_1, \ldots, \alpha_m$ are some numbers. Multiplying this equality by φ_j, with (11) taken into account, we get

$$\alpha_j(\varphi_j, \varphi_j) = 0.$$

Hence since $(\varphi_j, \varphi_j) > 0$, α_j is necessarily equal to zero for any j, that is, the system $\varphi_0, \varphi_1, \ldots, \varphi_m$ is linearly independent.

A Polynomial of Best Mean-square Approximation.

Definition. The function

$$\Phi_m(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_m\varphi_m(x), \tag{12}$$

where $c_0, c_1, \ldots, c_m$ are numerical coefficients, is called the *generalized polynomial* with respect to the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Let an arbitrary function f belong to E. The following problem is posed: find a polynomial (12) such that the distance $\rho(f, \Phi_m)$, which is also called the *mean-square deviation of the polynomial* Φ_m from the function f, is minimal. The polynomial Φ_m possessing the indicated property is called the *polynomial of best mean-square approximation* of the function f.

Let us show that *if the system of functions* $\varphi_0, \varphi_1, \ldots, \varphi_m$ *is linearly independent, then for any function* $f \in E$ *the polynomial of best mean-square approximation exists and is unique.* In accordance with (5) and (12),

$$\begin{aligned}\rho^2(f, \Phi_m) &= \|f - \Phi_m\|^2 = (f - \Phi_m, f - \Phi_m)\\ &= (f - c_0\varphi_0 - \ldots - c_m\varphi_m, f - c_0\varphi_0 - \ldots - c_m\varphi_m)\\ &= (f, f) + \sum_{j,k=0}^{m} c_j c_k(\varphi_j, \varphi_k) - 2\sum_{j=0}^{m} c_j(f, \varphi_j). \end{aligned}\tag{13}$$

Thus, the quantity $\rho^2(f, \Phi_m)$ is a quadratic form relative to the desired coefficients c_j of the polynomial (12). Since for any c_j, $j = 0, 1, \ldots, m$, $\rho^2(f, \Phi_m) \geqslant 0$, it is known from the theory of quadratic forms that the quadratic form (13) reaches its nonnegative minimum. Simultaneously with $\rho^2(f, \Phi_m)$, the distance $\rho(f, \Phi_m) = \sqrt{\rho^2(f, \Phi_m)}$ also reaches its minimum.

Equating the partial derivatives of the form (13) with respect to c_i, $i = 0, 1, \ldots, m$, to zero, reducing the coefficient equal to 2, and transposing the constant terms to the right, we arrive at the following system of linear algebraic equations:

$$\begin{aligned} &c_0(\varphi_0, \varphi_0) + c_1(\varphi_1, \varphi_0) + \ldots + c_m(\varphi_m, \varphi_0) = (f, \varphi_0),\\ &c_0(\varphi_0, \varphi_1) + c_1(\varphi_1, \varphi_1) + \ldots + c_m(\varphi_m, \varphi_1) = (f, \varphi_1),\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &c_0(\varphi_0, \varphi_m) + c_1(\varphi_1, \varphi_m) + \ldots + c_m(\varphi_m, \varphi_m) = (f, \varphi_m), \end{aligned}\tag{14}$$

called the *normal system*. By Lemma 1, the determinant of the system (14) which is the Gram determinant of the linearly independent system of

functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is not equal to zero. Therefore for any function $f \in E$ the system (14) has the unique solution $c_0, c_1, \ldots, c_m$ which corresponds to the unique stationary point of the quadratic form (13). This stationary point can be only the point of minimum, since the form (13) reaches its minimum.

Thus, if the system $\varphi_0, \varphi_1, \ldots, \varphi_m$ is linearly independent, then the coefficients of the unique polynomial of best mean-square approximation of the function $f \in E$ constructed with the aid of this system are found in the form of the solution of the normal system of linear algebraic equations (14).

In the space E_C of continuous functions with the scalar product (1) the distance $\rho(f, g)$, called the *mean-square distance*, in accordance with (5), acquires the form

$$\rho(f, g) = \sqrt{\frac{1}{b-a} \int_a^b (f(x) - g(x))^2 \, dx}, \tag{15}$$

and in the space E_{n+1} of functions defined on the discrete set $D = \{x_i\}_{i=0}^n$ with the scalar product (2) the mean-square distance is given by the formula

$$\rho(f, g) = \sqrt{\frac{1}{n+1} \sum_{i=0}^{n} (f(x_i) - g(x_i))^2}. \tag{16}$$

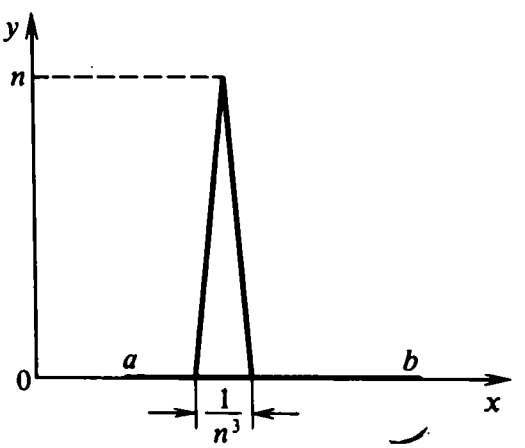


Fig. 5

It should be borne in mind that the proximity of two continuous functions by the distance (15), that is, in the sense of mean-square deviation, does not guarantee the smallness of their maximum deviation from each other. For instance, Fig. 5 shows two graphs: (1) of the function $g(x) \equiv 0$,

$x \in [a, b]$, and (2) of the function f, the latter differing from the former by a narrow tooth of altitude n and thickness at the base equal to $1/n^3$. Obviously, in this case,

$$\rho(f, g) = \sqrt{\frac{1}{b-a}\int_a^b (f(x) - g(x))^2\,dx}$$

$$= \sqrt{\frac{1}{b-a}\int_a^b f^2(x)\,dx} \leqslant \sqrt{\frac{1}{b-a}\,n^2\,\frac{1}{n^3}} = \frac{1}{\sqrt{(b-a)n}},$$

that is, by choosing n, we can make the mean-square distance $\rho(f, g)$ arbitrarily small and the quantity $\max\limits_{[a,b]} |f(x) - g(x)| = n$ arbitrarily large.

Mean-square Approximations by Algebraic Polynomials. In practice, we often use mean-square approximations of functions by algebraic polynomials, that is, instead of the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ we take the powers of x: $1, x, x^2, \ldots, x^m$. The system of these functions is linearly independent in E_C, that is, on the given interval $[a, b]$ for any m. In E_{n+1} it is linearly independent if $m \leqslant n$. This follows from the fact that under the condition $\sum\limits_{j=0}^{m} |\alpha_j| \neq 0$ the algebraic polynomial $\alpha_0 + \alpha_1 x + \ldots + \alpha_m x^m$ may vanish at not more than m points of the entire real axis. For $m \geqslant n + 1$ the system $1, x, x^2, \ldots, x^m$ is linearly dependent in E_{n+1}, since the polynomial (4.11) of $(n + 1)$th degree with the leading coefficient equal to unity vanishes on the entire set $D = \{x_i\}_{i=0}^{n}$.

Example. On the interval [0, 1] construct the polynomial of best mean-square approximation $\Phi_1(x) = c_0 + c_1 x$ for the function $f(x) = \sqrt{x}$.

Solution. $\varphi_0(x) = 1$, $\varphi_1(x) = x$, $(\varphi_0, \varphi_0) = \int_0^1 1^2\,dx = 1$,

$$(\varphi_1, \varphi_1) = \int_0^1 x^2\,dx = 1/3,\ (\varphi_0, \varphi_1) = (\varphi_1, \varphi_0) = \int_0^1 x\,dx = 1/2,$$

$$(f, \varphi_0) = \int_0^1 \sqrt{x}\,dx = 2/3,\ (f, \varphi_1) = \int_0^1 \sqrt{x}\,x\,dx = 2/5.$$

Consequently, the normal system of equations (14) is the following:

$$c_0 + \frac{1}{2} c_1 = \frac{2}{3},$$

$$\frac{1}{2} c_0 + \frac{1}{3} c_1 = \frac{2}{5}.$$

Hence, $c_0 = 4/15$, $c_1 = 4/5$, and $\Phi_1(x) = 4/15 + (4/5)x$. In this case, the mean-square deviation Φ_1 from f has the value

$$\rho(f, \Phi_1) = \sqrt{\int_0^1 \left(\sqrt{x} - \frac{4}{15} - \frac{4}{5} x\right)^2 dx} = \frac{\sqrt{2}}{30}.$$

When finding the algebraic polynomial of best mean-square approximation in E_C for the function f of a continuous argument on the interval $[a, b]$, we may encounter some difficulties in connection with the evaluation of the right-hand members of the equations (14), that is, the integrals $(f, \varphi_i) = \frac{1}{b-a} \int_a^b f(x) x^i dx$, $i = 0, 1, \ldots, m$, although the coefficients of the system, that is, the scalar products $(\varphi_j, \varphi_k) = \frac{1}{b-a} \times \int_a^b x^{j+k} dx$, are evaluated readily.

Therefore the method of least squares is used in the discrete variant, that is, in E_{n+1}. A set of points $\{x_i\}_{i=0}^n$ from the interval $[a, b]$ is given by experimental evaluations of the function f. It is desirable that the number of points exceeds the degree m of the polynomial at least by one and a half or two times. The points on the interval $[a, b]$ are arranged as uniformly as possible or are somewhat concentrated on the portion of the interval where it is important to obtain a more exact approximation of the function. In the discrete variant, the computation of the coefficients and right-hand sides of the normal system of equations (14) in terms of the scalar product (2) is not difficult.

The found algebraic polynomial of best mean-square approximation of the function f on the discrete set $\{x_i\}_{i=0}^n \subset [a, b]$ in sense of the distance (16) is usually taken for some approximating polynomial of the function f throughout the interval $[a, b]$.

Remark 1. If $m = n$, then the algebraic polynomial of degree n which is found in the discrete variant by the least-squares method coincides with the interpolation polynomial, since the deviation of the interpolation

polynomial from the given function f on the set of points $\{x_i\}_{i=0}^n$ in sense of the distance (16), by virtue of (4.1), is equal to zero and cannot be less than zero.

The mean-square approximation of a function by an algebraic polynomial is used when the function to be approximated is not sufficiently smooth and one fails to construct a suitable interpolation polynomial for it (see Secs. 4-7 and 9), spline (see Sec. 11) or a polynomial of uniform approximation (see Sec. 12), and also if the values of the function are known at a sufficiently large number of points, but with random errors. The effect of random errors on the approximating polynomial is considered in the next section.

Application of Orthogonal Polynomials. The solution of the normal system of equations (14) is found in a most simple way if the system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ is orthogonal, that is, satisfies the conditions (11). In this case, the matrix of the system (14) becomes diagonal, and the coefficients of the polynomial (12) of best mean-square approximation of the function f are

$$c_j = \frac{(f, \varphi_j)}{(\varphi_j, \varphi_j)}, \quad j = 0, 1, \ldots, m. \tag{17}$$

They are called the *Fourier coefficients* of the function f *with respect to the orthogonal system* $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Taking into account (4), (11), and (17), from (13) we find

$$\rho^2(f, \Phi_m) = \|f\|^2 - \sum_{j=0}^{m} c_j^2 \|\varphi_j\|^2, \tag{18}$$

where Φ_m is the polynomial of best mean-square approximation of the function f constructed with respect to the orthogonal system $\varphi_0, \varphi_1, \ldots, \varphi_m$, and c_j are its coefficients. Hence it is clearly seen that with an increase in m, that is, with new functions added to the orthogonal system $\varphi_0, \varphi_1, \ldots, \varphi_m$ (without changing the old ones), the quantity $\rho^2(f, \Phi_m)$, generally speaking, decreases (does not increase).

In the space E_{n+1} the orthogonal system $\varphi_0, \varphi_1, \ldots, \varphi_m$ may consist of not more than $n + 1$ functions, that is, $m \leqslant n$. If $m = n$ and this system is orthogonal, then it, being linearly independent, forms a basis in E_{n+1}. Then for any function $f \in E_{n+1}$ there is a polynomial $\Phi_n(x) = c_0\varphi_0(x) + \ldots + c_n\varphi_n(x)$ (a linear combination of functions $\varphi_j, j = 0, 1, \ldots, n$, with coefficients dependent on f) such that $\Phi_n(x) = f(x)$,

$x \in \{x_i\}_{i=0}^n$, and, consequently, $\rho(f, \Phi_m) = 0$. Obviously, the indicated polynomial coincides with the polynomial of best mean-square approximation of the function $f \in E_{n+1}$ for $m = n$.

On any interval $[a, b]$ there exists an infinite system of algebraic polynomials which is orthogonal in sense of the scalar product (1). Polynomials having the expression

$$X_n(x) = \frac{1}{n!2^n} \frac{d^n}{dx^n} (x^2 - 1)^n \tag{19}$$

are called *Legendre polynomials*. They possess the orthogonality property on the standard interval $[-1, 1]$:

$$(X_j, X_k) = \frac{1}{2} \int_{-1}^{1} X_j(x) X_k(x)\, dx = \begin{cases} 0, & j \neq k, \\ \dfrac{1}{2j+1}, & j = k, \end{cases} \tag{20}$$

which is verified by integration by parts. Obviously, $X_n(x)$ is an nth-degree algebraic polynomial with leading coefficient not equal to zero, since as a result of n-fold differentiation of the polynomial $(x^2 - 1)^n = x^{2n} - nx^{2n-2} + \ldots$, the degree is reduced exactly by n.

For Legendre polynomials the following reccurence formula holds true:

$$(n+1) X_{n+1}(x) - (2n+1) x X_n(x) + n X_{n-1}(x) = 0 \tag{21}$$

by which, taking into account that $X_0(x) = 1$ and $X_1(x) = x$, a Legendre polynomial of any degree may be found. In particular,

$$X_2(x) = \frac{1}{2}(3x^2 - 1), \quad X_3(x) = \frac{1}{2}(5x^3 - 3x),$$

$$X_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

The following lemma establishes the important property of Legendre polynomials, which will be used in Sec. 15.

Lemma 2. *All the roots of the Legendre polynomial* (19) *are real, simple, and found on the interval* $(-1, 1)$.

Proof. First of all let us note that the Legendre polynomial (19) is orthogonal to any algebraic polynomial $P_k(x)$ of degree $k < n$, that is,

$$\frac{1}{2} \int_{-1}^{1} P_k(x) X_n(x)\, dx = 0, \quad k < n. \tag{22}$$

This statement follows from (20) and the representation of the polynomial $P_k(x)$ in the form

$$P_k(x) = \alpha_0 X_0(x) + \alpha_1 X_1(x) + \ldots + \alpha_k X_k(x), \tag{23}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_k$ are quite definite numbers. In the expression (23), we first choose α_k in order that the leading coefficients coincide on the right and left, and then find α_{k-1}, and so forth.

Suppose that the polynomial $X_n(x)$ has only $k < n$ distinct real roots of odd multiplicity on the interval $(-1, 1)$. Let us denote these roots by $x_1, x_2, \ldots, x_k$ and specify the polynomial of degree k:

$$P_k(x) = \begin{cases} (x - x_1)(x - x_2) \ldots (x - x_k), & k > 0, \\ 1, & k = 0. \end{cases} \tag{24}$$

Obviously, the polynomial of degree $n + k$, being the product $P_k(x)\, X_n(x)$, does not change sign on the interval $(-1, 1)$ where it may have roots only of even multiplicity and is not identically equal to zero. Therefore

$$\int_{-1}^{1} P_k(x)\, X_n(x)\, dx \neq 0,$$

which contradicts the equality (22). Consequently, the Legendre polynomial (19) of degree n has exactly n simple (single) roots on the interval $(-1, 1)$.

Lemma 2 has been proved.

Example. Approximate the function $f(x) = |x|$ on the interval $[-1, 1]$ by a fourth-degree algebraic polynomial with the aid of the method of least squares.

Solution. We present the desired polynomial $\Phi_4(x)$ in terms of Legendre polynomials:

$$\Phi_4(x) = c_0 X_0(x) + c_1 X_1(x) + \ldots + c_4 X_4(x).$$

In accordance with (1), (17), and (20), we have

$$c_j = \frac{2j+1}{2} \int_{-1}^{1} |x|\, X_j(x)\, dx.$$

In particular, $c_0 = 1/2$, $c_1 = 0$, $c_2 = 5/8$, $c_3 = 0$, and $c_4 = -3/16$. Thus,

$$\Phi_4(x) = \frac{1}{2} + \frac{5}{8 \cdot 2}(3x^2 - 1) - \frac{3}{16 \cdot 8}(35x^4 - 30x^2 + 3)$$

$$= \frac{15}{128}(-7x^4 + 14x^2 + 1).$$

Since

$$\|f\|^2 = (f, f) = \frac{1}{2}\int_{-1}^{1} |x|^2\, dx = \frac{1}{3},$$

according to (20) and (18), where $\varphi_j = X_j$, we obtain the following value of the mean-square deviation of the polynomial $\Phi_4(x)$ from $f(x) = |x|$ on $[-1, 1]$:

$$\rho(f, \Phi_4) = \sqrt{\frac{1}{3} - \left(\frac{1}{2}\right)^2 - \left(\frac{5}{8}\right)^2 \frac{1}{5} - \left(-\frac{3}{16}\right)^2 \frac{1}{9}} = \frac{\sqrt{3}}{48}.$$

It is possible to prove that the Legendre polynomials $X_0(x)$, $X_1(x)$, . . ., $X_m(x)$, . . . form a complete system of functions, that is, whatever the function $f \in C[-1, 1]$, its polynomials $\Phi_m(x)$ of the best mean-square approximation which are constructed in terms of Legendre polynomials converge to the function f in the mean, that is, the mean-square distance

$$\rho(f, \Phi_m) = \sqrt{\frac{1}{2}\int_{-1}^{1} (f(x) - \Phi_m(x))^2\, dx}$$

tends to zero as $m \to \infty$. But convergence in the mean does not guarantee that at each point x of the interval $[-1, 1]$ there exists the limit $\lim_{m \to \infty} \Phi_m(x)$ equal to $f(x)$.

There are algebraic polynomials, orthogonal on a discrete set of points, that is, in sense of the scalar product (2).

The polynomial $x^{(p)} = x(x - 1) \ldots (x - (p - 1))$ is called the *factorial polynomial* of degree p, $x^{(0)} = 1$.

For every natural n the *system of Chebyshev polynomials* is defined:

$$P_{mn}(x) = \sum_{j=0}^{m} (-1)^j \frac{C_m^j C_{m+j}^j x^{(j)}}{n^{(j)}}, \tag{25}$$

$m = 0, 1, \ldots, n$, *which are orthogonal on the discrete set of integral points* $\{x_i\}_{i=0}^{n}$, $x_i = i$, that is,

$$(P_{mn}, P_{jn}) = \begin{cases} 0, & j \neq m, \\ \dfrac{(m + n + 1)^{(m+1)}}{n^{(m)}(2m + 1)(n + 1)}, & j = m, \end{cases} \tag{26}$$

where $(f, g) = \dfrac{1}{n + 1} \sum_{i=0}^{n} f(i)\, g(i)$.

In particular, we have

$$P_{0n}(x) = 1,$$

$$P_{1n}(x) = 1 - 2\frac{x}{n}, \tag{27}$$

$$P_{2n}(x) = 1 - 6\frac{x}{n} + 6\frac{x(x - 1)}{n(n - 1)}.$$

Mean-square Approximations by Trigonometric Polynomials. The function

$$\Phi_m(x) = a_0 + \sum_{p=1}^{m} (a_p \cos px + b_p \sin px), \tag{28}$$

where a_p and b_p are arbitrary numerical coefficients, is called the *trigonometric polynomial of order m.*

It is natural to use trigonometric polynomials for approximating periodic functions with period 2π. In the theory of Fourier series it is established that the coefficients of the trigonometric polynomial of best mean-square approximation of a continuous 2π-periodic function f for which the quantity

$$\rho(f, \Phi_m) = \sqrt{\frac{1}{2\pi} \int_0^{2\pi} (f(x) - \Phi_m(x))^2\, dx}$$

attains the minimum value are given by the formulas*

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,dx,\ a_p = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos px\,dx,$$
$$b_p = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin px\,dx,\ p > 0, \tag{29}$$

that is, are Fourier coefficients. We are not going to dwell on this question in more detail, noting only that, in practice, the computation of the Fourier coefficients (29) which are given in the form of integrals may involve considerable difficulties.

Consider the problem of finding the trigonometric polynomial of best mean-square approximation on a discrete set of points. Let n and m be natural numbers, $m \leqslant n/2$,

$$D = \{x_i\}_{i=0}^{n}, \tag{30}$$

where $x_i = 2\pi i/(n+1)$, $i = 0, 1, \ldots, n$. Let us set

$$\varphi_0(x) = 1,\ \varphi_p(x) = \cos px,\ \psi_p(x) = \sin px, \tag{31}$$
$$p = 1, 2, \ldots, m.$$

It is possible to prove that the trigonometric system of functions (31) is orthogonal on the discrete set (30) in sense of the scalar product (2), namely,

$$(\varphi_j, \psi_k) = 0, j = 0, 1, \ldots, m, k = 1, 2, \ldots, m,$$
$$(\varphi_r, \varphi_s) = 0, r \neq s, (\psi_t, \psi_q) = 0, t \neq q, \tag{32}$$
$$(\varphi_0, \varphi_0) = 1, (\varphi_p, \varphi_p) = (\psi_p, \psi_p) = \frac{1}{2},$$
$$p = 1, 2, \ldots, m \leqslant \frac{n}{2}.$$

Therefore the coefficients of the trigonometric polynomial of best mean-square approximation of the function f on the discrete set (30)

---

* See *Differential Equations. Multiple Integrals. Series. Theory of Functions of a Complex Variable* by Ya. S. Bugrov and S. M. Nikolsky, Secs. 4.3 and 4.9 (in English—Mir Publishers, Moscow, 1983).

which has the form

$$\Phi_m(x) = \alpha_0 + \sum_{p=1}^{m} (\alpha_p \cos px + \beta_p \sin px) \tag{33}$$

are found, according to (17), (2), and (32), by the formulas

$$\begin{aligned} \alpha_0 &= \frac{1}{n+1} \sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right), \\ \alpha_p &= \frac{2}{n+1} \sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right) \cos p \frac{2\pi i}{n+1}, \\ \beta_p &= \frac{2}{n+1} \sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right) \sin p \frac{2\pi i}{n+1}, \quad p = 1, 2, \ldots, m. \end{aligned} \tag{34}$$

In this case, analogously to (18), we obtain

$$\rho(f, \Phi_m) = \sqrt{\frac{1}{n+1} \sum_{i=0}^{n} (f(x_i) - \Phi_m(x_i))^2}$$

$$= \sqrt{\|f\|^2 - \alpha_0^2 - \frac{1}{2} \sum_{p=1}^{m} (\alpha_p^2 + \beta_p^2)},$$

where $\|f\|^2 = (f, f) = \dfrac{1}{n+1} \displaystyle\sum_{i=0}^{n} f^2\left(\frac{2\pi i}{n+1}\right)$.

For an even n and $m = n/2$, the trigonometric system (31) which is orthogonal on D consists of $2m + 1 = n + 1$ functions and forms a basis in E_{n+1}. Therefore the trigonometric polynomial (33) of best mean-square approximation of the function f in E_{n+1} whose coefficients are computed by the formulas (34) for the indicated n and m is an interpolation polynomial, that is, $\Phi_{n/2}(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$, and $\rho(f, \Phi_{n/2}) = 0$.

Remark 2. As it has already been noted above, it is not always useful, for instance, when the values of a function are known with errors, to approximate the function by an interpolation polynomial. These errors also

bring distortions in the polynomial which is found by the method of least squares. These distortions become stronger with an increase in m, that is, in the order (or degree) of the polynomial, while the interpolation polynomial is better the larger the possible m for the given n. This question is discussed in detail in the next section.

Sec. 14. Analysis of Errors of Mean-square Approximations. "Smoothing" of Observations

This section studies the errors of the mean-square approximations obtained by the method of least squares in the discrete variant. We consider here the effect of random errors in function values (errors of observation results) and analyze the error of the method due to the fact that the approximated function does not belong to the class of polynomials by which the approximation is performed. Separate consideration of the random error and the error of the method is allowed, since the operator for constructing the polynomial of best mean-square approximation is linear. In conclusion, the procedure of smoothing the observations is explained.

Let, as in Sec. 13, E_{n+1} be Euclidean space of functions defined on the finite set of points $\{x_i\}_{i=0}^{n} \subset [a, b]$ with the scalar product (13.2). Suppose that given in E_{n+1} is an orthogonal system of functions $\varphi_0, \varphi_1, \ldots, \varphi_m$, that is, a function satisfying the conditions (13.11), where $m \leqslant n^*$. Using this system of functions, we construct the approximating polynomial.

Suppose that instead of the values $f(x_i)$ of the function f its approximate values are known:

$$f_i^* = f(x_i) + \eta_i, \; i = 0, 1, \ldots, n.$$

Here, η_i are errors of observations which are independent random quantities with zero mean value and variance independent of i, that is, in particular,

$$\mathbf{M}[\eta_i] = 0, \; i = 0, 1, \ldots, n, \tag{1}$$

$$\mathbf{M}[\eta_i \eta_j] = \begin{cases} 0, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{2}$$

where $\mathbf{M}[\xi]$ is the expectation of the random quantity ξ. Then, instead of the coefficients (13.17) of the polynomial of best mean-square approx-

---

* Since orthogonal functions are linearly independent, their number $m + 1$ cannot exceed the dimension of the space E_{n+1} equal to $n + 1$.

imation of the function f which has the form (13.12), we shall find the coefficients

$$c_j^* = \frac{(f^*, \varphi_j)}{(\varphi_j, \varphi_j)} = \frac{(f + \eta, \varphi_j)}{(\varphi_j, \varphi_j)} = \frac{(f, \varphi_j)}{(\varphi_j, \varphi_j)} + \frac{(\eta, \varphi_j)}{(\varphi_j, \varphi_j)} = c_j + \gamma_j, \tag{3}$$

where

$$\gamma_j = \frac{(\eta, \varphi_j)}{(\varphi_j, \varphi_j)}, \tag{4}$$

γ_j being the random error of the jth coefficient.

Hence, with (13.2) and (1) in mind, we obtain

$$\mathbf{M}[\gamma_j] = \mathbf{M}\left[\frac{(\eta, \varphi_j)}{(\varphi_j, \varphi_j)}\right] = \frac{1}{(\varphi_j, \varphi_j)} \mathbf{M}[(\eta, \varphi_j)]$$

$$= \frac{1}{(\varphi_j, \varphi_j)} \mathbf{M}\left[\frac{1}{n+1} \sum_{i=0}^{n} \eta_i \varphi_j(x_i)\right]$$

$$= \frac{1}{(n+1)(\varphi_j, \varphi_j)} \sum_{i=0}^{n} \varphi_j(x_i) \mathbf{M}[\eta_i] = 0. \tag{5}$$

Thus, the expectation of the error of each coefficient of the approximating polynomial is equal to zero.

Taking into consideration (4) and (13.2), we find

$$\mathbf{M}[\gamma_j \gamma_k]$$

$$= \frac{1}{(n+1)^2(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} \mathbf{M}\left[\sum_{i=0}^{n} \eta_i \varphi_j(x_i) \sum_{q=0}^{n} \eta_q \varphi_k(x_q)\right]$$

$$= \frac{1}{(n+1)^2(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} \sum_{i=0}^{n} \sum_{q=0}^{n} \varphi_j(x_i)\, \varphi_k(x_q)\, \mathbf{M}[\eta_i, \eta_q].$$

By virtue of (2), in a double sum only the terms corresponding to $q = i$ may differ from zero. Hence, with (13.11) taken into account, we get

$$\mathbf{M}[\gamma_j\gamma_k] = \frac{1}{(n+1)(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)}\left(\frac{1}{n+1}\sum_{i=0}^{n}\varphi_j(x_i)\,\varphi_k(x_i)\,\sigma^2\right)$$

$$= \frac{(\varphi_j, \varphi_k)\sigma^2}{(n+1)(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} = \begin{cases} 0, & k \neq j, \\ \dfrac{\sigma^2}{(n+1)(\varphi_j, \varphi_j)}, & k = j, \end{cases} \tag{6}$$

that is, the errors of different coefficients are uncorrelated, and the variance of the error of the jth coefficient has the form

$$\mathbf{D}[\gamma_j] = \frac{\sigma^2}{(n+1)(\varphi_j, \varphi_j)}, \tag{7}$$

where σ^2 is the variance of observation errors.

On the basis of (3), we have

$$\Phi_m^*(x) = \sum_{j=0}^{m} c_j^*\varphi_j(x) = \sum_{j=0}^{m}(c_j + \gamma_j)\varphi_j(x) = \Phi_m(x) + \Gamma_m(x),$$

where

$$\Gamma_m(x) = \gamma_0\varphi_0(x) + \gamma_1\varphi_1(x) + \ldots + \gamma_m\varphi_m(x), \tag{8}$$

$\Gamma_m(x)$ is the random error of the approximating polynomial $\Phi_m^*(x)$ found by the method of least squares, and $\Phi_m(x)$ is the polynomial of best mean-square approximation of the function f with random errors absent. In (8), the argument x can attain a value from the finite set $\{x_i\}_{i=0}^n$ and any other values for which the functions $\varphi_0, \varphi_1, \ldots, \varphi_m$ are defined. In particular, if instead of the indicated functions we take the Chebyshev polynomials (13.25) which are orthogonal on a discrete set, then x can run throughout the real axis.

Taking into consideration (5)-(8) and that, according to the well-known theorems of the theory of probability, the expectation of the sum of random quantities is equal to the sum of their expectations, and the variance of the sum of uncorrelated random quantities with some coefficients is equal to the sum of their variances multiplied by the squared

coefficients, we obtain

$$\mathbf{M}[\Gamma_m(x)] = \varphi_0(x)\,\mathbf{M}[\gamma_0] + \ldots + \varphi_m(x)\,\mathbf{M}[\gamma_m] = 0, \tag{9}$$

$$\mathbf{D}[\Gamma_m(x)] = \frac{\sigma^2}{n+1}\sum_{j=0}^{m}\frac{\varphi_j^2(x)}{(\varphi_j, \varphi_j)}. \tag{10}$$

Thus, the expectation of the random error $\Gamma_m(x)$ of the polynomial of best mean-square approximation is equal to zero, and its variance for any fixed x, generally speaking, grows with an increase in m, since this causes the appearance of new nonnegative terms in (10).

Remark 1. If the observation errors η_i are distributed according to the normal law, then since according to (4) and (8), the error $\Gamma_m(x)$ in the approximating polynomial depends linearly on observation errors, it is also distributed normally for any fixed x.

The Periodic Case. Let $[a, b] = [0, 2\pi]$ and let the values of a 2π-periodic function f be known at the points $x_i = 2\pi i/(n+1)$, $i = 0, 1, \ldots, n$. The polynomial of best mean-square approximation of the function f on the discrete set $\{x_i\}_{i=0}^n$ which is constructed according to the orthogonal trigonometric system of functions (13.31) has the form (13.33), its coefficients being found by the formulas (13.34).

If for the construction of an approximating polynomial we use the values of the function f with errors η_i which are random quantities with zero mean possessing Property (2), then in accordance with (10), (13.31), and (13.32), we have for $m \leqslant n/2$:

$$\mathbf{D}[\Gamma_m(x)] = \frac{\sigma^2}{n+1}\left(\frac{\varphi_0^2(x)}{(\varphi_0, \varphi_0)} + \sum_{p=1}^{m}\left(\frac{\varphi_p^2(x)}{(\varphi_p, \varphi_p)} + \frac{\psi_p^2(x)}{(\psi_p, \psi_p)}\right)\right)$$

$$= \frac{\sigma^2}{n+1}\left(1 + 2\sum_{p=1}^{m}(\cos^2 px + \sin^2 px)\right) = \frac{2m+1}{n+1}\sigma^2. \tag{11}$$

Thus, the variance of the error $\Gamma_m(x)$ of the approximating trigonometric polynomial of order m grows proportionally to m and is independent of x, and according to (9), $\mathbf{M}[\Gamma_m(x)] = 0$. The mean-square value of the error $\Gamma_m(x)$ will be denoted by $\sigma_m(x)$:

$$\sigma_m(x) = \sqrt{\mathbf{D}[\Gamma_m(x)]} = \sigma\sqrt{\frac{2m+1}{n+1}}, \quad m \leqslant \frac{n}{2}, \tag{12}$$

where σ is the mean-square value of observation errors.

Let us now dwell on the distortions of the Fourier coefficients due to discreteness. Let us assume that the continuous periodic function f with period 2π is expanded into the uniformly convergent Fourier series

$$f(x) = a_0 + \sum_{p=1}^{\infty}{}' (a_p \cos px + b_p \sin px), \tag{13}$$

where a_0, a_p, and b_p are the Fourier coefficients (13.29). Then it is possible to prove that if, for instance, n is even, then the coefficients (13.34) of the polynomial (13.33) of best mean-square approximation of the function f are expressed in terms of the Fourier coefficients (13.29) of the function f (in the absence of observation errors) in the following way:

$$\begin{aligned} \alpha_0 &= a_0 + a_n + a_{2n} + \ldots, \\ \alpha_p &= a_p + a_{n-p} + a_{n+p} + a_{2n-p} + a_{2n+p} + \ldots, \\ \beta_p &= b_p - b_{n-p} + b_{n+p} - b_{2n-p} + b_{2n+p} - \ldots, \end{aligned} \tag{14}$$

$$1 \leqslant p \leqslant m \leqslant \frac{n}{2}.$$

If the Fourier coefficients a_p and b_p decrease rapidly, then α_p and β_p are close to a_p and b_p for small p, while the relative distortions usually increase with an increase in p.

The coefficients (13.34) may also be regarded as the approximations by the rectangular quadrature formulas (see Sec. 15) of the integrals (13.29) (strictly speaking, the integrals equal to them within the range from π/n to $2\pi + \pi/n$). In the supposition that the periodic function f under consideration is twice continuously differentiable throughout the real axis, from the known formula $(fg)'' = f''g + 2f'g' + fg''$ there readily follow the inequalities:

$$\begin{aligned} \max_{(-\infty, +\infty)} |(f(x) \cos px)''| &\leqslant M_2 + 2pM_1 + p^2M_0, \\ \max_{(-\infty, +\infty)} |(f(x) \sin px)''| &\leqslant M_2 + 2pM_1 + p^2M_0, \end{aligned} \tag{15}$$

where $p = 0, 1, \ldots, m$,

$$M_q = \max_{(-\infty, +\infty)} |f^{(q)}(x)|. \tag{16}$$

On the basis of (15), (16), and the estimate (15.31) of the error of the rectangular formula, there follow the estimates

$$|a_0 - \alpha_0| \leqslant \frac{\pi^2}{6n^2} M_2,$$

$$|a_p - \alpha_p| \leqslant \frac{\pi^2}{3n^2} (M_2 + 2pM_1 + p^2M_0), \qquad (17)$$

$$|b_p - \beta_p| \leqslant \frac{\pi^2}{3n^2} (M_2 + 2pM_1 + p^2M_0),$$

where $p = 1, 2, \ldots, m \leqslant n/2$.

It remains to estimate the remainder of the Fourier series, that is, the error due to the replacement of the Fourier series (13) by its partial sum of order m. Under the condition that a 2π-periodic function f is continuously differentiable q times on the real axis, applying q times integration by parts in the integrals (13.29) and taking into account that for a natural p

$$\int_0^{2\pi} |\cos px|\, dx = \int_0^{2\pi} |\sin px|\, dx = 4,$$

we obtain the estimates

$$|a_p|, |b_p| \leqslant \frac{4}{\pi} \frac{M_q}{p^q}, p \geqslant 1,$$

where M_q is the quantity (16). Hence, since $|\cos px| + |\sin px| \leqslant \sqrt{2}$, for $q \geqslant 2$ there follows the inequality

$$\left| \sum_{p=m+1}^{\infty} (a_p \cos px + b_p \sin px) \right| \leqslant \frac{4\sqrt{2}}{\pi} M_q \int_m^{\infty} \frac{dp}{p^q} = \frac{4\sqrt{2}}{\pi} \frac{M_q}{q-1} \frac{1}{m^{q-1}}. \qquad (18)$$

Thus, we have considered three components of the error due to approximating a periodic function by a trigonometric polynomial using the method of least squares, namely, the random error, then the error of the initial Fourier coefficients due to discreteness, and finally, the remainder of the Fourier series.

Let us draw some conclusions. An increase in n, that is, in the number of points x_i on the interval $[0, 2\pi]$ involves a reduction of the variance (11) of the random error and favours the reduction of distortions of the Fourier coefficients due to discreteness (in any case, the right-hand sides of the estimates (17) tend to zero as $n \to \infty$). For instance, if $n = 27$ and $m = 3$, then according to (12), the mean-square error of the approximating polynomial will be half the mean-square observation error. In this, in particular, there consists the smoothing effect of the method of least squares.

For a fixed n, when choosing $m \leqslant n/2$, one should observe the compromise between the random error growing, according to (12), with an increase in m and also the error due to the distortions of the Fourier coefficients whose number is proportional to m, and the error due to the rejection of the remainder of the Fourier series whose estimate (18) becomes better with an increase in m.

The Nonperiodic Case. Consider the mean-square approximations of the functions defined on the set $\{i\}_{i=0}^{n}$ of integral points by the orthogonal Chebyshev polynomials (13.25). The approximating polynomial of degree $m \leqslant n$ for the function f found by the method of least squares has the form

$$\Phi_m(x) = \sum_{j=0}^{m} c_j P_{jn}(x), \tag{19}$$

where, according to (13.26) and (13.17),

$$c_j = \frac{(f, P_{jn})}{(P_{jn}, P_{jn})} \tag{20}$$

and $(f, g) = \dfrac{1}{n+1} \sum_{i=0}^{n} f(i)\, g(i).$

If the values of the function f at the points of the set $\{i\}_{i=0}^{n}$ are known with the random errors η_i possessing Properties (1) and (2), then, according to (9) and (10),

$$\mathbf{M}[\Gamma_m(x)] = 0, \ \mathbf{D}[\Gamma_m(x)] = \frac{\sigma^2}{n+1} \sum_{j=0}^{m} \frac{P_{jn}^2(x)}{(P_{jn}, P_{jn})}, \tag{21}$$

where $\Gamma_m(x)$ is the random error of the approximating polynomial. In

particular, when approximating by a first-degree polynomial, that is, for $m = 1$, on the basis of (21), (13.26), and (13.27), we have

$$\mathbf{D}[\Gamma_1(x)] = \frac{\sigma^2}{n+1}\left(1 + \frac{12(x - n/2)^2}{n(n+2)}\right). \tag{22}$$

Thus, with an increase in $m (m \leqslant n)$, that is, in the degree of the approximating polynomial (19), for fixed n and x the variance of the random error, generally speaking, increases. In contrast to the periodic case, the variance of the random error is not constant, namely: in the middle of the observation interval it is considerably less than at its ends. For $m = 1$ it is readily seen from the formula (22). In addition, in order to illustrate what was said, let us give the table of the values of the variance of the error $\Gamma_m(x)$ for $n = 10$ and $\sigma^2 = 1$.

Table 3

Values of $\mathbf{D}[\Gamma_m(x)]$

m	0	1	2	3	4	...	10
$x = 0, 10$	0.09	0.32	0.58	0.79	0.91	...	1.00
$x = 5$	0.09	0.09	0.21	0.21	0.33	...	1.00

Remark 2. If observations are carried out with a positive step $h \neq 1$, that is, at the points $x = 0, h, 2h, \ldots, nh$, then the Chebyshev polynomials $P_{mn}(x)$ are replaced everywhere by $P_{mn}(x/h)$ and, consequently, in the right-hand member of the formula (22) x is replaced by x/h. In this case, the scalar product is specified in the following way:

$$(f, g) = \frac{1}{n+1}\sum_{i=0}^{n} f(ih)g(ih).$$

Consider the error of the method. Suppose there are no observation errors, but the function f is not an algebraic polynomial of degree m. Let us assume that the observation interval $[0, nh]$ is not large. Then if $f \in C_{m+1}[0, nh]$, in first approximation we may regard that the function f is an algebraic polynomial of degree $m + 1$ with leading coefficient equal to $f^{(m+1)}(x_{\text{mean}})/(m+1)!$, where $x_{\text{mean}} = nh/2$. Let us assume for definiteness that the function f is a polynomial of degree $m + 1$ with the indicated leading coefficient, and $m < n$. Then the method of least

squares will lead to the identities

$$f(x) \equiv \Phi_{m+1}(x) \equiv c_{m+1}P_{m+1,n}\left(\frac{x}{h}\right) + \Phi_m(x), \tag{23}$$

where $\Phi_m(x)$ is expressed by (19) where the right-hand member $P_{jn}(x)$ is replaced by $P_{jn}(x/h)$.

The first identity holds true because any polynomial of degree $m + 1$ is exactly reconstructed by the method of least squares if the degree of the approximating polynomial is equal to $m + 1$, and $m + 1 \leqslant n$. The second identity expresses the fact that $\Phi_{m+1}(x)$ is obtained, according to (19), from $\Phi_m(x)$ by adding one term. Thus, the error of approximating the polynomial $f(x)$ of degree $m + 1$ by the method of least squares in terms of the polynomial of degree m acquires the form

$$f(x) - \Phi_m(x) = c_{m+1}P_{m+1,n}\left(\frac{x}{h}\right), \tag{24}$$

where the coefficient c_{m+1} can be found either by the formula (20) with Remark 2 taken into account or from the condition of the coincidence of the leading coefficients in both sides of the identities (23). With (13.25) in mind, using the second method, we find

$$c_{m+1} = (-1)^{m+1}h^{m+1}\frac{(m+1)!}{(2m+2)!}n^{(m+1)}f^{(m+1)}(x_{\text{mean}}). \tag{25}$$

In particular, if

$$f(x) = \frac{ax^2}{2} + bx + c \quad (f'' = a > 0),$$

$m = 1$, then by virtue of (24), (25), and (13.27),

$$f(x) - \Phi_1(x) = ah^2\frac{n(n-1)}{12}\left(1 - \frac{6}{n}\frac{x}{h} + \frac{6}{n(n-1)}\frac{x}{h}\left(\frac{x}{h} - 1\right)\right),$$

and for $n \gg 1$

$$f(x) - \Phi_1(x) \approx a\frac{l^2}{12}\left(1 - 6\frac{x}{l}\left(1 - \frac{x}{l}\right)\right),$$

where $l = nh$ is the length of the observation interval. The graph of the polynomial on the right is shown in Fig. 6.

In the above-considered example, for $m = 1$ the error of the method, the same as the random error, in the middle of the observation interval is less than at its ends. An analogous phenomenon takes place for other m.

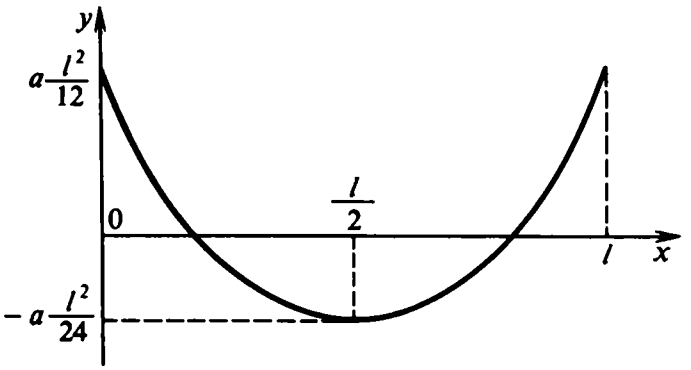


Fig. 6

On the whole, after having considered the random error and the error of the method in the nonperiodic case, we draw the following conclusions. Instants of observations (points of measurement) are advisable to be chosen so that the most interesting section turns out to be nearer to the middle of the observation interval where both the mean-square random error of an approximating polynomial and the error of the method are usually smaller. When carrying out observations and choosing the degree m of an approximating polynomial, a compromise between the random error and the error of the method should be observed. The mean-square value of the random error grows with an increase in m ($m \leqslant n$), and decreases with an increase in n. The error of the method usually grows with an increase in nh, and decreases with an increase in m and a decrease in h for a fixed n.

"Smoothing" the Observations. From the formulas (19) and (20) it is seen that the value of the approximating polynomial at each fixed point x is expressed linearly in terms of the observed values. Let $n = 6$ and $m = 3$. Let us denote by $\overline{f}_i$ the value of the polynomial $\Phi_3(x)$ at the midpoint of the observation interval, and by $f_{i-3}, f_{i-2}, f_{i-1}, f_i, f_{i+1}, f_{i+2}$, and f_{i+3} the used observed values whose number is equal to $n + 1 = 7$. Then

$$\overline{f}_i = \frac{1}{21}(-2f_{i-3} + 3f_{i-2} + 6f_{i-1} + 7f_i + 6f_{i+1} + 3f_{i+2} - 2f_{i+3}). \quad (26)$$

Smoothing consists in that the table of observations f_i with a constant step is replaced by the table of the values $\overline{f}_i$ computed, say, by the formula (26). This table is supposed to be long, that is, to contain more than $n + 1$ values. At the edges of the table smoothing is realized by other formulas

obtained from the expression for the polynomial $\Phi_3(x)$ constructed by the corresponding extreme knots.

In addition to the formula (26), smoothing can be effected by the formulas obtained from the expressions of the polynomials (19) for other values of n and m. Given below are some more smoothing formulas:

$n = 2k$ and $m = 1$:

$$\bar{f}_i = \frac{1}{n+1}\left(f_i + \sum_{j=1}^{k}(f_{i-j} + f_{i+j})\right);$$

$n = 4$ and $m = 3$:

$$\bar{f}_i = \frac{1}{35}(-3f_{i-2} + 12f_{i-1} + 17f_i + 12f_{i+1} - 3f_{i+2});$$

$n = 8$ and $m = 5$:

$$\bar{f}_i = \frac{1}{429}(15f_{i-4} - 55f_{i-3} + 30f_{i-2} + 135f_{i-1} + 179f_i + 135f_{i+1} + 30f_{i+2} - 55f_{i+3} + 15f_{i+4}).$$

A curve constructed by smoothed values becomes smoother by virtue of the following reasons:

(1) If the observed values contain random errors with properties similar to (1) and (2), then the variance of the error in the smoothed values decreases. In particular, when making use of the formula (26), we have

$$\bar{\sigma}^2 = \frac{1}{3}\sigma^2 \quad \left(\bar{\sigma} = \frac{1}{\sqrt{3}}\sigma\right),$$

where $\bar{\sigma}$ is the mean-square value of the error of smoothed values, and σ is the mean-square error of observations.

(2) The errors in smoothed values $\bar{f}_i$ become strongly correlated. When using the formula (26) and making the same allowances for the observation errors as indicated above, we have the following correlation coefficients of the smoothed values:

$$r_{i,i\pm1} = 0.73, \quad r_{i,i\pm2} = 0.37, \quad r_{i,i\pm3} = 0.05,$$
$$r_{i,i\pm4} = -0.10, \quad r_{i,i\pm5} = -0.08, \quad r_{i,i\pm6} = 0.03.$$

(3) If the function f is not an algebraic polynomial of degree m, then its graph turns out to be "blurred".

Repeated smoothing is not recommended.

Chapter 2

NUMERICAL INTEGRATION

In practice, we seldom succeed in finding the exact value of a definite integral or in integrating an ordinary differential equation. For instance, the integral $\int_1^2 \frac{dx}{\ln x}$ cannot be expressed in elementary functions, and the equation $u' = e^{-(x^2 + u^2)}$ cannot be integrated.

Section 15 treats in detail quadrature formulas (the rectangular formula, trapezoidal formula, and Simpson's formula) widely used for approximately evaluating definite integrals, and also Gaussian quadrature formulas which turn out to be exact for algebraic polynomials of higher degree. Section 16 justifies Runge's rule for a practical error estimate, in particular, quadrature formulas, and introduces Richardson's method for making approximate solutions more precise. Section 17 is dedicated to the Monte Carlo method for the approximate evaluation of definite integrals. Section 18 describes the approximate numerical methods of solving Cauchy's problem for the ordinary differential equation $u' = f(x, u)$, analyzes in detail the simplest Euler method, introduces the Runge-Kutta methods, and gives a brief description of the Adams method. The Runge-Kutta methods and the Adams method are compared in this section.

Sec. 15. Quadrature Formulas

Let us first formulate a theorem from integral calculus which will be used later.

Theorem 1 (generalized mean-value theorem). *Let $f, g \in C[a, b]$, and $g(x) \geqslant 0$ on $[a, b]$. Then there is a point $\xi \in [a, b]$ such that*

$$\int_a^b f(x)\, g(x)\, dx = f(\xi) \int_a^b g(x)\, dx.$$

Proof. Let us set

$$M = \max_{[a, b]} f(x), \quad m = \min_{[a, b]} f(x). \tag{1}$$

Then since $g(x) \geqslant 0$, we have

$$mg(x) \leqslant f(x)\, g(x) \leqslant Mg(x), \quad x \in [a, b],$$

and, consequently,

$$m \int_a^b g(x)\, dx \leqslant \int_a^b f(x)\, g(x)\, dx \leqslant M \int_a^b g(x)\, dx.$$

Hence the inequalities $\int_a^b g(x)\, dx \geqslant 0$ and $M \geqslant m$ lead to the existence of a number c which satisfies the inequalities

$$m \leqslant c \leqslant M \tag{2}$$

such that

$$\int_a^b f(x)\, g(x)\, dx = c \int_a^b g(x)\, dx. \tag{3}$$

According to the theorem on intermediate values of a continuous function, by virtue of (1) and (2), there is a point $\xi \in [a, b]$, at which $f(\xi) = c$, which together with the equality (3) proves Theorem 1.

Let us introduce the notion of a quadrature formula. Let there be given the definite integral

$$I = \int_a^b f(x)\, dx \tag{4}$$

of a function f which is continuous on the interval $[a, b]$. The approximate equality

$$\int_a^b f(x)\, dx \approx \sum_{j=1}^{n} q_j f(x_j), \tag{5}$$

where q_j are certain numbers, and x_j are certain points of the interval $[a, b]$, is called a *quadrature formula* defined by *weights* q_j and *nodes* x_j.

A quadrature formula is said to be exact for polynomials of degree m if, on replacing f by an arbitrary algebraic polynomial of degree m, the approximate equality (5) becomes exact.

Let us first consider the simplest quadrature formulas.

The Rectangular Formula. Suppose that $f \in C_2[-h/2, h/2]$, $h > 0$. We then set approximately

$$\int_{-h/2}^{h/2} f(x)\,dx \approx hf_0, \tag{6}$$

where $f_0 = f(0)$, that is, the area of the curvilinear trapezoid bounded from above by the graph of the function f is approximated by the area of the hatched rectangle (Fig. 7) whose altitude is equal to the value of f at

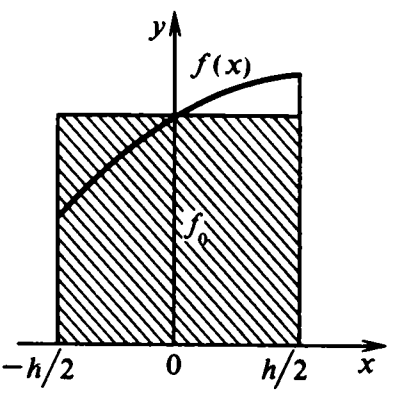


Fig. 7

the midpoint of the trapezoid's base. We then find the remainder, that is, the error of the formula (6). Let

$$F(x) = \int_0^x f(t)\,dt, \tag{7}$$

$F_{\pm 1/2} = F(\pm h/2)$. Since $F(0) = 0$, $F'(0) = f_0$, $F''(0) = f_0' = f'(0)$, and $F'''(x) = f''(x)$, according to Taylor's formula with its remainder in Lagrange's form, we have

$$F_{\pm 1/2} = \pm \frac{h}{2} f_0 + \frac{h^2}{8} f_0' \pm \frac{h^3}{48} f''(\xi_\pm), \tag{8}$$

where ξ_- and ξ_+ are some points, $-h/2 < \xi_- < \xi_+ < h/2$.

The function (7) is an antiderivative for $f(x)$. Therefore for the integral on the left-hand side of the approximate equality (6), from the

Newton-Leibniz formula (with (8) in mind), we obtain the following expression:

$$\int_{-h/2}^{h/2} f(x)\,dx = F_{1/2} - F_{-1/2} = hf_0 + \frac{h^3}{24}\frac{f''(\xi_-) + f''(\xi_+)}{2}.$$

Hence with the aid of Lemma 1.10, we obtain the *rectangular formula with its remainder:*

$$\int_{-h/2}^{h/2} f(x)\,dx = hf_0 + \frac{h^3}{24}f''(\xi),\ |\xi| \leqslant \frac{h}{2}. \tag{9}$$

The Trapezoidal Formula. Let $f \in C_2[0, h]$. We set

$$\int_0^h f(x)\,dx \approx h\frac{f_0 + f_1}{2}, \tag{10}$$

where $f_0 = f(0)$ and $f_1 = f(h)$, i.e. the integral $\int_0^h f(x)\,dx$ is approximately replaced by the area of the hatched trapezoid shown in Fig. 8. Let us express f_1 and $F_1 = F(h)$, where F is the function (7), by Taylor's formula with its remainder in integral form:

$$f_1 = f_0 + hf_0' + \int_0^h (h - t)f''(t)\,dt, \tag{11}$$

$$F_1 = F(0) + hF'(0) + \frac{h^2}{2}F''(0) + \frac{1}{2}\int_0^h (h - t)^2 F'''(t)\,dt$$

$$= hf_0 + \frac{h^2}{2}f_0' + \frac{1}{2}\int_0^h (h - t)^2 f''(t)\,dt. \tag{12}$$

According to (11), we have

$$h\frac{f_0}{2} = h\frac{f_1}{2} - \frac{h^2}{2}f_0' - \frac{h}{2}\int_0^h (h - t)f''(t)\,dt. \tag{13}$$

Isolating the term $hf_0/2$ on the right-hand side of (12) and replacing it by the expression (13), on account of the fact that $F_1 = \int_0^h f(x)\,dx$, we find

$$\int_0^h f(x)\,dx = h\frac{f_0 + f_1}{2} - \frac{1}{2}\int_0^h (h - t)\,tf''(t)\,dt.$$

Since $(h - t)t \geqslant 0$, $t \in [0, h]$, by Theorem 1, we have

$$\int_0^h (h - t)\,tf''(t)\,dt = f''(\xi)\int_0^h (h - t)t\,dt = \frac{h^3}{6}f''(\xi),$$

where $\xi \in [0, h]$ is a point. Thus, we have arrived at the *trapezoidal formula with its remainder*

$$\int_0^h f(x)\,dx = h\frac{f_0 + f_1}{2} - \frac{h^3}{12}f''(\xi), \quad \xi \in [0, h]. \tag{14}$$

Simpson's Formula. Suppose that $f \in C_4[-h, h]$. We replace approximately the integral $\int_{-h}^h f(x)\,dx$ by the area of the hatched curvilinear

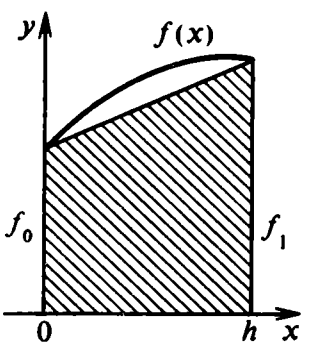


Fig. 8

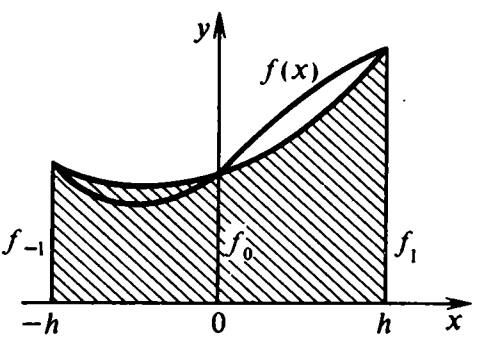


Fig. 9

trapezoid (Fig. 9) bounded from above by a parabola passing through the points $(-h, f_{-1})$, $(0, f_0)$, and (h, f_1), where $f_i = f(ih)$. This parabola is given by the equation

$$y = f_0 + \frac{f_1 - f_{-1}}{2h}x + \frac{f_{-1} - 2f_0 + f_1}{2h^2}x^2,$$

which is readily verified by setting, in turn, x equal to $-h$, 0, and h. Hence we easily find

$$\int_{-h}^{h} y\,dx = \frac{h}{3}(f_{-1} + 4f_0 + f_1).$$

Thus, *Simpson's formula,* which is also called the *parabolic formula,* has the form

$$\int_{-h}^{h} f(x)\,dx \approx \frac{h}{3}(f_{-1} + 4f_0 + f_1). \tag{15}$$

Let us set $F_{\pm 1} = F(\pm h)$, where F is the function (7). Since $F(0) = 0$ and $F^{(k)}(x) = f^{(k-1)}(x)$, $1 \leqslant k \leqslant 5$, by Taylor's formula with its remainder in integral form, we have

$$F_{\pm 1} = \pm hf_0 + \frac{h^2}{2}f_0' \pm \frac{h^3}{6}f_0'' + \frac{h^4}{24}f_0''' \pm \frac{1}{24}\int_0^h (h-t)^4 f^{(4)}(\pm t)\,dt,$$

$$f_{\pm 1} = f_0 \pm hf_0' + \frac{h^2}{2}f_0'' \pm \frac{h^3}{6}f_0''' + \frac{1}{6}\int_0^h (h-t)^3 f^{(4)}(\pm t)\,dt.$$

Hence we obtain

$$F_1 - F_{-1} - \frac{h}{3}(f_{-1} + 4f_0 + f_1) = -\frac{1}{24}\int_0^h (h-t)^3 \left(\frac{h}{3} + t\right)(f^{(4)}(t) + f^{(4)}(-t))\,dt \tag{16}$$

(the rest of the terms cancel out). Since $(h-t)^3\,(h/3 + t) \geqslant 0$, $t \in [0, h]$, applying Theorem 1 to the integral (16) and then Lemma 10.1 to the result obtained, we find

$$-\frac{1}{24}\int_0^h (h-t)^3\left(\frac{h}{3}+t\right)(f^{(4)}(t)+f^{(4)}(-t))\,dt$$

$$=-\frac{1}{12}\,\frac{f^{(4)}(\eta)+f^{(4)}(-\eta)}{2}\int_0^h (h-t)^3\left(\frac{h}{3}+t\right)dt$$

$$=-\frac{h^5}{90}f^{(4)}(\xi), \quad (17)$$

where $\eta \in [0, h]$ and $\xi \in [-h, h]$ are some points.

Taking into account that $F_1 - F_{-1} = \int_{-h}^{h} f(x)\,dx$, from (16) and (17) we arrive at *Simpson's formula with its remainder:*

$$\int_{-h}^{h} f(x)\,dx = \frac{h}{3}(f_{-1}+4f_0+f_1) - \frac{h^5}{90}f^{(4)}(\xi). \quad (18)$$

The quadrature formulas considered above ((6), (10), and (15)) are called *canonical.*

Composite Quadrature Formulas. In practice, if it is required to evaluate approximately the integral (1), then the given interval $[a, b]$ is divided into N equal subintervals, one of the canonical quadrature formulas is applied on each of the subintervals, and the results obtained are summed. The quadrature formula thus constructed on the interval $[a, b]$ is said to be *composite.* When applying the rectangular and trapezoidal formulas, it is convenient to take subintervals of length h, and when using Simpson's formula, of length $2h$.

Let us dwell on the use of the rectangular formula in more detail. Let $f \in C_2[a, b]$. We denote the subintervals by $[x_i, x_{i+1}]$, where $x_i = a + ih$, $i = 0, 1, \ldots, N-1$, $x_N = b$, and $h = (b-a)/N$. In accordance with (6), we set

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx hf_{i+1/2}, \quad (19)$$

where $f_{i+1/2} = f(a + (i + 1/2)h)$ is the value of f at the midpoint of the subinterval $[x_i, x_{i+1}]$. Here the following equality is valid which is

analogous to (9):

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = hf_{i+1/2} + \frac{h^3}{24} f''(\xi_i), \tag{20}$$

where $\xi_i \in [x_i, x_{i+1}]$ is some point. Summing over all subintervals of the approximation (19) leads to the *composite rectangular formula:*

$$\int_a^b f(x)\,dx \approx h(f_{1/2} + f_{3/2} + \ldots + f_{N-1/2}), \tag{21}$$

and summing the equalities (20), bearing in mind that by Lemma 10.1

$$\sum_{i=0}^{N-1} f''(\xi_i) = N\left(\frac{1}{N}\sum_{i=0}^{N-1} f''(\xi_i)\right) = Nf''(\xi) = \frac{b-a}{h} f''(\xi),$$

where ξ is a point of the interval $[a, b]$, yields the *composite rectangular formula with its remainder:*

$$\int_a^b f(x)\,dx = h(f_{1/2} + f_{3/2} + \ldots + f_{N-1/2}) + h^2 \frac{b-a}{24} f''(\xi). \tag{22}$$

Just in the same way, under the condition that $f \in C_2[a, b]$ and using the formulas (10) and (14), we obtain the *composite trapezoidal formula:*

$$\int_a^b f(x)\,dx \approx h\left(\frac{f_0}{2} + f_1 + f_2 + \ldots + f_{N-1} + \frac{f_N}{2}\right) \tag{23}$$

and the corresponding *formula with its remainder:*

$$\int_a^b f(x)\,dx = h\left(\frac{f_0}{2} + f_1 + f_2 + \ldots + f_{N-1} + \frac{f_N}{2}\right) - h^2 \frac{b-a}{12} f''(\xi), \tag{24}$$

where $f_i = f(a + ih)$, $h = (b-a)/N$, and $\xi \in [a, b]$ is some point.

Let now $h = (b - a)/2N$ and, as usually, $x_j = a + jh$, $f_j = f(x_j)$. We rewrite Simpson's canonical formula (15) in connection with subinterval $[x_{2i}, x_{2i+2}]$ of length $2h$:

$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,dx \approx \frac{h}{3}(f_{2i} + 4f_{2i+1} + f_{2i+2}).$$

Summing both sides of this relationship over i from 0 to $N - 1$, we obtain *Simpson's composite formula:*

$$\int_a^b f(x)\,dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 4f_{2N-1} + f_{2N})$$

$$= \frac{h}{3}\left(f_0 + f_{2N} + 4\sum_{i=1}^{N} f_{2i-1} + 2\sum_{i=1}^{N-1} f_{2i}\right). \quad (25)$$

The corresponding *formula with its remainder,* which is obtained by summing equalities of the form (18) over the subintervals $[x_{2i}, x_{2i+2}]$ provided that $f \in C_4[a, b]$, is:

$$\int_a^b f(x)\,dx$$

$$= \frac{h}{3}\left(f_0 + f_{2N} + 4\sum_{i=1}^{N} f_{2i-1} + 2\sum_{i=1}^{N-1} f_{2i}\right) - h^4\frac{b-a}{180}f^{(4)}(\xi), \quad (26)$$

where $f_i = f(a + ih)$, $h = (b - a)/2N$, and $\xi \in [a, b]$.

For the sake of brevity, let us introduce the notation

$$I_h^{\text{rect}} = h\sum_{i=0}^{N-1} f_{i+1/2}, \quad I_h^{\text{trap}} = h\left(\frac{f_0 + f_N}{2} + \sum_{i=1}^{N-1} f_i\right), \quad (27)$$

where $h = (b - a)/N$ and $f_\mu = f(a + \mu h)$, and let us also set

$$I_h^{\text{Simp}} = \frac{h}{3}\left(f_0 + f_{2N} + 4\sum_{i=1}^{N} f_{2i-1} + 2\sum_{i=1}^{N-1} f_{2i}\right), \quad (28)$$

where $h = (b - a)/2N$ and $f_i = f(a + ih)$.

The approximate equalities

$$I = \int_a^b f(x)\,dx \approx I_h^{\text{rect}}, \quad I = \int_a^b f(x)\,dx \approx I_h^{\text{trap}}, \tag{29}$$

$$I = \int_a^b f(x)\,dx \approx I_h^{\text{Simp}} \tag{30}$$

will be called the *rectangular formula,* the *trapezoidal formula,* and *Simpson's formula,* respectively, with the words 'composite' and 'quadrature' omitted. In the notation I_h^{rect}, I_h^{trap}, and I_h^{Simp} h instead of N is chosen as the parameter, since h, being the step used with the values of the function f in the formulas under consideration, is more characteristic.

From the expressions for the remainder in (22), (24), and (26) it is seen that the rectangular and trapezoidal formulas (29) are exact for polynomials of first degree, that is, for linear functions, whereas Simpson's formula (30) is exact for polynomials of third degree (for them the remainder is equal to zero). The error of the formulas (29) has the second order of accuracy with respect to h (it is, *a fortiori,* not better if f'' is continuous on $[a, b]$ and does not vanish), while Simpson's formula is a formula of the fourth order of accuracy if f is smooth. Therefore for the functions of class $C_4[a, b]$ for a small h Simpson's formula usually yields a higher accuracy as compared with the formulas (29).

When evaluating the integral (4), by virtue of (22) and (26), the error of the rectangular formula and Simpson's formula satisfies the inequalities

$$|I - I_h^{\text{rect}}| \leqslant h^2 \frac{b-a}{24} \max_{[a, b]} |f''(x)|, \tag{31}$$

$$|I - I_h^{\text{Simp}}| \leqslant h^4 \frac{b-a}{180} \max_{[a, b]} |f^{(4)}(x)|. \tag{32}$$

There is a similar inequality for the error of the trapezoidal formula.

Along with the estimates of an error from above, estimates from below are also useful. In particular, for the error of the rectangular formula the lower estimate, which follows from (22), is:

$$|I - I_h^{\text{rect}}| \geqslant h^2 \frac{b-a}{24} \min_{[a, b]} |f''(x)|. \tag{33}$$

As an example, let us analyze the errors of the quadrature formulas for the integral

$$I = \int_0^{1/2} e^{-x^2}\,dx$$

which is not expressible in terms of elementary functions and is frequently used in applications.

We have $f(x) = e^{-x^2}$, $f''(x) = (4x^2 - 2)e^{-x^2}$, $f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2}$, $e^{-1/4} \leqslant |f''(x)| \leqslant 2$, $|f^{(4)}(x)| \leqslant 12$ on $[0, 1/2]$. Hence for $h = 0.05$, by (31)-(33), we get

$$0.4 \cdot 10^{-4} \leqslant |I - I_h^{\text{rect}}| \leqslant 0.11 \cdot 10^{-3}, \quad |I - I_h^{\text{Simp}}| \leqslant 0.21 \cdot 10^{-6},$$

that is, the upper estimate of the error of Simpson's formula is considerably less than the lower estimate of the error of the rectangular formula.

When integrating smooth functions, the rectangular and trapezoidal formulas, taken separately, yield Simpson's formula. But when applied pairwise, they possess a valuable property, namely, if f'' does not change sign on $[a, b]$, then the formulas (29) yield two-sided approximations for the integral (4), since, according to (22) and (24), their remainders have opposite signs. In the foregoing example $f''(x) < 0$ and $x \in [0, 1/2]$. Therefore

$$I_h^{\text{trap}} < I < I_h^{\text{rect}}.$$

In this situation, it is natural to set $I \approx (I_h^{\text{rect}} + I_h^{\text{trap}})/2$. Then

$$\left| I - \frac{I_h^{\text{rect}} + I_h^{\text{trap}}}{2} \right| < \frac{I_h^{\text{rect}} - I_h^{\text{trap}}}{2},$$

that is, the error is estimated by means of approximate values of the integral.

The rectangular and trapezoidal formulas yield two-sided approximations of the integral (4) not only when f'' does not change sign. We shall return to this point in Sec. 16.

The Gaussian Quadrature Formula. There is a direct connection between the highest power of polynomials for which the quadrature formula is exact and the order of accuracy of the quadrature formula, say, with respect to the step h with which the values of the integrand are used. For example, the rectangular and trapezoidal formulas (29) are exact for polynomials of first degree and possess the second order of accuracy with

respect to h. And Simpson's formula (25), being exact for polynomials of third degree, has the fourth order of accuracy for $f \in C_4[a, b]$.

In connection with the foregoing there arises a natural problem of finding among all the quadrature formulas (5) with n nodal points a quadrature formula with such an arrangement of nodes x_j on the interval $[a, b]$ and such weights q_j that it is exact for algebraic polynomials of maximum degree. It is clear that this degree is less than $2n$, since for any choice of the nodes x_j and weights q_j, $j = 1, 2, \ldots, n$ the polynomial $P_{2n}(x) = (x - x_1)^2 (x - x_2)^2 \ldots (x - x_n)^2$ of degree $2n$ possesses such a property that

$$\int_a^b P_{2n}(x)\,dx > 0, \quad \text{but} \sum_{j=1}^{n} q_j P_{2n}(x_j) = 0.$$

Consider the standard interval $[a, b] = [-1, 1]$. Let for the time being $x_j \in [-1, 1]$, $j = 1, 2, \ldots, n$, be arbitrary pairwise nonintersecting nodes. Then if we take

$$q_j = \int_{-1}^{1} p_{n-1,j}(x)\,dx, \quad j = 1, 2, \ldots, n, \tag{34}$$

where

$$p_{n-1,j}(x) = \frac{(x - x_1)\ldots(x - x_{j-1})(x - x_{j+1})\ldots(x - x_n)}{(x_j - x_1)\ldots(x_j - x_{j-1})(x_j - x_{j+1})\ldots(x_j - x_n)},$$

$p_{n-1,j}(x)$ are the Lagrangian coefficients for the interpolation polynomial of degree $n - 1$, then the quadrature formula will be exact in any case for polynomials of degree $n - 1$. Indeed, any algebraic polynomial $P_{n-1}(x)$ of degree $n - 1$ (see Remark 4.1) is exactly reconstructed from the interpolation polynomial $L_{n-1}(x)$ of the same degree. Therefore

$$\int_{-1}^{1} P_{n-1}(x)\,dx$$

$$= \int_{-1}^{1} L_{n-1}(x)\,dx = \int_{-1}^{1} \sum_{j=1}^{n} p_{n-1,j}(x) P_{n-1}(x_j)\,dx$$

$$= \sum_{j=1}^{n} \left(\int_{-1}^{1} p_{n-1,j}(x)\,dx \right) P_{n-1}(x_j) = \sum_{j=1}^{n} q_j P_{n-1}(x_j).$$

Let us now take the roots of the Legendre polynomial $X_n(x)$ (see (13.19)) as the nodes x_j, $j = 1, 2, \ldots, n$. According to Lemma 13.2, these roots are located in the interval $(-1, 1)$. We find the weights q_j by the formula (34). Let us show that the quadrature formula

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{j=1}^{n} q_j f(x_j) \tag{35}$$

with the properly chosen nodes and weights is exact for polynomials of degree $2n - 1$, thus, for the maximum degree.

Let $P_{2n-1}(x)$ be an arbitrary algebraic polynomial of degree $2n - 1$. We represent it in the form

$$P_{2n-1}(x) = U_{n-1}(x)X_n(x) + V_{n-1}(x),$$

where $U_{n-1}(x)$ and $V_{n-1}(x)$ are polynomials of degree $n - 1$ (the quotient and remainder from the division of $P_{2n-1}(x)$ by the Legendre polynomial $X_n(x)$).

By virtue of the orthogonality of the Legendre polynomial $X_n(x)$ to any polynomial of degree $n - 1$ (see (13.22)), we have

$$\int_{-1}^{1} P_{2n-1}(x)\,dx = \int_{-1}^{1} U_{n-1}(x)\,X_n(x)\,dx + \int_{-1}^{1} V_{n-1}(x)\,dx$$

$$= \int_{-1}^{1} V_{n-1}(x)\,dx. \tag{36}$$

Since $X_n(x_j) = 0, j = 1, 2, \ldots, n$, we have

$$\sum_{j=1}^{n} q_j P_{2n-1}(x_j) = \sum_{j=1}^{n} q_j U_{n-1}(x_j)\,X_n(x_j) + \sum_{j=1}^{n} q_j V_{n-1}(x_j)$$

$$= \sum_{j=1}^{n} q_j V_{n-1}(x_j). \tag{37}$$

Since the quadrature formula (35) is automatically exact for polynomials of degree $n - 1$, we have

$$\int_{-1}^{1} V_{n-1}(x)\,dx = \sum_{j=1}^{n} q_j V_{n-1}(x_j),$$

combining this with (36) and (37), we arrive at the desired equality

$$\int_{-1}^{1} P_{2n-1}(x)\,dx = \sum_{j=1}^{n}{}' q_j P_{2n-1}(x_j).$$

The quadrature formula (35) with n indicated nodes and weights which is exact for polynomials of degree $2n - 1$ is called the *Gaussian formula.* It is possible to prove that the nodes of the Gaussian formula, that is, the roots of the Legendre polynomial $X_n(x)$, are situated symmetrically with respect to the point $x = 0$, and the weights q_j are positive and coincide at symmetrical nodes for any n. Here we present the numerical expressions of nonnegative nodes x_j and weights q_j of the Gaussian formula for $n = 1, 2, 3, 4$ to ten decimal places.

n	x_1 q_1	x_2 q_2	n	x_1 q_1	x_2 q_2
1	0.00000 00000 2.00000 00000		3	0.00000 00000 0.88888 88888	0.77459 66692 0.55555 55555
2	0.57735 02692 1.00000 00000		4	0.33998 10436 0.65214 51549	0.86113 63116 0.34785 48451

The Gaussian formula is somewhat inconvenient, since its nodes and weights are irrational in the general case. But this is compensated for by the fact that this formula is exact for polynomials of higher degree than other quadrature formulas with n nodes, except for the particular case $n = 1$, when the Gaussian formula coincides with the canonical rectangular formula (6), where $h = 2$.

With the aid of the Gaussian quadrature formula (35) with n nodes on the standard interval $[-1, 1]$, which will be called *canonical,* we can construct the composite Gaussian quadrature formula on an arbitrary interval $[a, b]$. To this end, let us break the interval $[a, b]$ into N equal subintervals $[x_k^*, x_{k+1}^*]$, where $x_k^* = a + k(b - a)/N$, $k = 0, 1, \ldots, N - 1$, and $x_N^* = b$. On each subinterval $[x_k^*, x_{k+1}^*]$ we specify n nodes:

$$x_{kj} = \frac{x_k^* + x_{k+1}^*}{2} + x_j \frac{b - a}{2N}, \quad j = 1, 2, \ldots, n, \tag{38}$$

where x_j are the nodes of the Gaussian canonical formula (35). The arrangement of the nodes x_{kj}, $j = 1, 2, \ldots, n$, on the subinterval $[x_k^*,$

$x_{k+1}^*]$ is geometrically similar to that of the nodes x_j of the Gaussian canonical formula on the interval $[-1, 1]$. By virtue of this, the quadrature formula

$$\int_{x_k^*}^{x_{k+1}^*} f(x)\,dx \approx \frac{b-a}{2N} \sum_{j=1}^{n} q_j f(x_{kj}), \tag{39}$$

where q_j are the weights of the Gaussian canonical formula (35), is exact for polynomials of degree $2n-1$. Summing the relationship (39) over k from 0 to $N-1$, we obtain the *composite Gaussian quadrature formula:*

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2N} \sum_{j=1}^{n} q_j \sum_{k=0}^{N-1} f(x_{kj}) \tag{40}$$

which is also exact for polynomials of degree $2n-1$. If $f \in C_{2n}[a, b]$, its remainder R_n has the form

$$R_n = \frac{(b-a)^{2n+1}}{N^{2n}} \frac{(n!)^4}{((2n)!)^3(2n+1)} f^{(2n)}(\xi). \tag{41}$$

In particular,

$$R_2 = \frac{(b-a)^5}{4320N^4} f^{(4)}(\xi), \quad R_3 = \frac{(b-a)^7}{2{,}016{,}000N^6} f^{(6)}(\xi). \tag{42}$$

Let us compare the composite Gaussian quadrature formula (40) for $n = 2$ with composite Simpson's formula (25). Since in Simpson's formula (25) h and N are connected by the relationship $h = (b-a)/2N$, the remainder of the formula (25) which is denoted by R_{Simp} can be represented, according to (26), in the form

$$R_{\text{Simp}} = -h^4 \frac{b-a}{180} f^{(4)}(\xi) = -\frac{(b-a)^5}{2880N^4} f^{(4)}(\xi).$$

Thus, the remainder (42) of the composite Gaussian formula which corresponds to $n = 2$ has the numerical coefficient which is 3/2 the remainder of Simpson's formula by absolute value. Both of these formulas are exact for polynomials of third degree. When using Simpson's formula (25), it is required to compute $2N+1$ values of the given function, and when applying the formula (40) with $n = 2$, the number of the values of the function used is equal to $2N$. And still preference should be given to Simpson's formula (25) in which the nodes are spaced with constant step h on the interval $[a, b]$.

The Gaussian quadrature formula, the composite formula included, is expedient to be used for $n > 2$ for approximately evaluating integrals of highly smooth functions.

Sec. 16. Runge's Rule for a Practical Error Estimate

Taking Account of Excess Smoothness of the Integrand. Suppose that $f \in C_4[a, b]$ and that the rectangular formula (15.21) is applied for approximately evaluating the integral (15.4). According to (15.22), its error expressed in terms of f'' is the quantity $O(h^2)$. In the general case, by virtue of the relationship (15.22), the upper estimate of the error of the rectangular formula (15.31) cannot be improved by the order with respect to h, since there are the additional derivatives f''' and $f^{(4)}$. For instance, if $\min_{[a, b]} |f''(x)| > 0$, the lower estimate of the error (15.33) which has the exact second order with respect to h is also valid along with (15.31).

When using the relationship (15.22), the existence of the "superfluous" derivatives f''' and $f^{(4)}$ in f yields nothing (these derivatives are not taken into account in the relationship). However, if $f \in C_4[a, b]$, another relationship can be obtained, which is, in a sense, more interesting.

Under the condition that $f \in C_4[-h/2, h/2]$, we have, along with (15.8)

$$F_{\pm 1/2} = \pm\frac{h}{2} f_0 + \frac{h^2}{8} f_0'' \pm \frac{h^3}{48} f_0''' + \frac{h^4}{384} f_0'''' \pm \frac{h^5}{3840} f^{(4)}(\eta_\pm),$$

$-h/2 < \eta_- < \eta_+ < h/2$. Hence, analogously to (15.9), we obtain

$$\int_{-h/2}^{h/2} f(x)\,dx = F_{1/2} - F_{-1/2}$$

$$= hf_0 + \frac{h^3}{24} f_0'' + \frac{h^5}{1920} f^{(4)}(\eta), \quad |\eta| \leqslant \frac{h}{2}. \quad (1)$$

Thus, the following equality holds true:

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = hf_{i+1/2} + \frac{h^3}{24} f''_{i+1/2} + \frac{h^5}{1920} f^{(4)}(\eta_i), \quad (2)$$

where $x_i = a + ih$, $h = (b - a)/N$, $f^{(k)}_{i+1/2} = f^{(k)}(a + (i + 1/2)h)$ is the value of $f^{(k)}$ at the midpoint of the subinterval $[x_i, x_{i+1}]$ and

$\eta_i \in [x_i, x_{i+1}]$ is some point, $i = 0, 1, \ldots, N-1$. Summing the equality (2) with respect to i from 0 to $N-1$, analogously to (15.22), we obtain

$$I = I_h^{\text{rect}} + \frac{h^2}{24}\left(h \sum_{i=0}^{N-1} f''_{i+1/2}\right) + h^4 \frac{b-a}{1920} f^{(4)}(\eta), \tag{3}$$

where I is the integral (15.4), I_h^{rect} has the form (15.27), and $\eta \in [a, b]$ is some point.

According to the equality (15.22) with $f'' \in C_2[a, b]$ substituted for f, we have

$$\int_a^b f''(x)\,dx = h \sum_{i=0}^{N-1} f''_{i+1/2} + h^2 \frac{b-a}{24} f^{(4)}(\xi), \tag{4}$$

where $\xi \in [a, b]$ is some point. From (3) and (4) we find

$$I = I_h^{\text{rect}} + ch^2 + O(h^4), \tag{5}$$

where c is a constant independent of h:

$$c = \frac{1}{24} \int_a^b f''(x)\,dx. \tag{6}$$

The quantity ch^2 in (5) is called the *principal part of the error* of the rectangular formula. It may happen that $c = 0$. Then the principal part of the error is equal to zero, and the error of the rectangular formula is a quantity of order h^4. But usually $c \neq 0$.

If $f \in C_4[a, b]$, then the following relationship is also valid:

$$I = I_h^{\text{trap}} + c_1 h^2 + O(h^4), \tag{7}$$

where I_h^{trap} is defined in (15.27) and is an approximate value of the integral (15.4) which is found by the trapezoidal formula with step h, $c_1 = -(1/12)\int_a^b f''(x)\,dx$ is independent of h.

Remark 1. From the relationships (5) and (7) it follows, in particular, that if $\int_a^b f''(x)\,dx \neq 0$ and $f \in C_4[a, b]$, then with a sufficiently small h

the rectangular and trapezoidal formulas (15.29) yield two-sided approximations of the integral (15.4) even if f'' does not preserve sign on $[a, b]$.

Under the condition that $f \in C_6[a, b]$ analogously to (5), it is possible to obtain the following relationship:

$$I = I_h^{\text{Simp}} + ch^4 + O(h^6), \tag{8}$$

where I is the integral (15.4), I_h^{Simp} is its approximate value found by Simpson's formula, that is, the value (15.28), and c is some constant independent of h.

Runge's Rule. Let z be the unknown exact value of some quantity, and let z_h be its known approximate value (approximate solution) dependent on a positive parameter h which can attain arbitrarily small values.

Suppose we have established the following relationship:

$$z = z_h + ch^k + O(h^{k+m}), \tag{9}$$

where c is an unknown constant independent of h, and $k > 0, m > 0$ are known numbers. Then

$$z = z_{h/2} + c\left(\frac{h}{2}\right)^k + O(h^{k+m}). \tag{10}$$

Subtracting the equalities (9) and (10), we find

$$z_{h/2} - z_h = c\left(\frac{h}{2}\right)^k (2^k - 1) + O(h^{k+m}).$$

Hence

$$c\left(\frac{h}{2}\right)^k = \frac{z_{h/2} - z_h}{2^k - 1} + O(h^{k+m}) \tag{11}$$

and, consequently, according to (10), with an accuracy to $O(h^{k+m})$, we have

$$z - z_{h/2} \approx \frac{z_{h/2} - z_h}{2^k - 1}, \tag{12}$$

where z_h and $z_{h/2}$ are known quantities.

If $c \neq 0$, then by virtue of (11), the right-hand side of (12) has exactly the kth order with respect to h and differs from the principal part of the error, i.e. from $c(h/2)^k$, by a quantity of a higher order with respect to h.

The computation of an approximate estimate of the error by the formula (12) with the condition (9) fulfilled is called *Runge's rule*.

Remark 2. In practice, the condition $c \neq 0$ is confirmed by the following inequality:

$$\left| 2^k \frac{z_h - z_{h/2}}{z_{2h} - z_h} - 1 \right| < 0.1. \tag{13}$$

And only in this case is Runge's rule recommended to be applied. The inequality (13) cannot be fulfilled owing to the following reasons: (a) h is large, then the effect of the negligible term $O(h^{k+m})$ occurs; (b) h is too small, then the round-off errors may affect the calculations on a computer; (c) $c = 0$ or is close to zero.

Richardson's Refinement of an Approximate Solution. Subtracting the equality (9) from the equality (10) which is multiplied by 2^k, we get

$$z(2^k - 1) = 2^k z_{h/2} - z_h + O(h^{k+m}).$$

Hence

$$z = z_h^* + O(h^{k+m}), \tag{14}$$

where

$$z_h^* = \frac{2^k z_{h/2} - z_h}{2^k - 1}. \tag{15}$$

The number z_h^* is called the *Richardson refined* (or *extrapolated*) *approximate value* of the quantity z.

According to (14), $z - z_h^* = O(h^{k+m})$, and at the same time, if in (9) $c \neq 0$, then $z - z_{h/2}$ has exactly the kth order with respect to h.

Thus, with the condition (9) imposed, where $c \neq 0$, in terms of the approximate solutions z_h and $z_{h/2}$, it is possible, firstly, to estimate approximately the error of $z_{h/2}$ by Runge's rule, that is, by the formula (12), and, secondly, to compute by the formula (15) the approximate solution z_h^* having the error of a higher order than $z_{h/2}$ with respect to h.

Application to Quadrature Formulas. Let us denote by I_h the approximate value of the integral I (15.4) found by one of the three formulas (15.29),(15.30) and unite (5), (7), and (8) in one relationship:

$$I = I_h + ch^k + O(h^{k+2}), \tag{16}$$

where c is independent of h; $k = 2$ for the triangular and trapezoidal formulas, and $k = 4$ for Simpson's formula. It is supposed that $f \in C_{k+2}[a, b]$.

The relationship (16) is of the form (9). Therefore, according to the aforegoing, on computing I_{2h}, I_h, $I_{h/2}$ and making sure that the quantity

$2^k(I_h - I_{h/2})/(I_{2h} - I_h)$ is close to unity, it is possible to approximately estimate the error $I - I_{h/2}$ using Runge's rule:

$$I - I_{h/2} \approx \frac{I_{h/2} - I_h}{2^k - 1}.$$

In addition, it is possible to find the Richardson refined value of the integral I:

$$I_h^* = \frac{2^k I_{h/2} - I_h}{2^k - 1}$$

with the error $I - I_h^* = O(h^{k+2})$.

Remark 3. The trapezoidal and Simpson's formulas are more convenient to be applied here, since when passing from h to $h/2$ all the values of the function f computed earlier may be used.

Remark 4. Under the condition that $f \in C_{2n+2}[a, b]$, the relationship (16) holds, where I_h is the approximate value of the integral (15.4) found by the composite Gaussian formula (15.40), $k = 2n$, and $h = 1/N$.

Other Applications. Runge's rule for error estimate and the Richardson refined approximate solution may be applied when solving other problems in terms of approximate methods if the relationship (9) is determined, where c is independent of h. In particular, this is possible in the numerical differentiation of sufficiently smooth functions. Let $f \in C_6[x_0 - h_0, x_0 + h_0]$, $h_0 > 0$. Then, using Taylor's formula with its remainder expressible in terms of $f^{(6)}$ to represent the values of the function $f_{\pm 1} = f(x_0 \pm h)$, $0 < h \leqslant h_0$, we obtain, analogously to (10.3), the relationship

$$f_0'' = f_{0h}'' - \frac{h^2}{12} f_0^{(4)} - \frac{h^4}{360} f^{(6)}(\xi), \tag{17}$$

where $f_0^{(k)} = f^{(k)}(x_0)$, $|\xi - x_0| < h$, f_{0h}'' is an approximate value of f_0'' equal to the right-hand side of (10.7). The relationship (17) has the form (9), where $k = m = 2, c = -f_0^{(4)}/12$, and is independent of h. Thus, the principal condition of the applicability of Runge's rule is fulfilled.

Sec. 17. The Monte Carlo Method

The Monte Carlo method consists in that a random quantity ξ is given whose expectation is equal to the desired quantity z, that is,

$$\mathbf{M}[\xi] = z; \tag{1}$$

a series of n independent trials of the random quantity ξ is realized: $\xi_1, \xi_2, \ldots, \xi_n$ and it is supposed that

$$\bar{\xi}_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} \approx z. \tag{2}$$

By virtue of (1), for any natural n

$$\mathbf{M}[\bar{\xi}_n] = \mathbf{M}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}[\xi_i] = \frac{nz}{n} = z. \tag{3}$$

If the variance $\mathbf{D}[\xi] = \sigma^2$ is finite, then

$$\mathbf{D}[\bar{\xi}_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}[\xi_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \tag{4}$$

the distribution of the random quantity $\bar{\xi}_n$ being asymptotically normal by virtue of the central limit theorem. Therefore for a sufficiently large n (practically, for $n > 10$), by (3), (4), and the known "three sigmas" rule, we have

$$p\left(|z - \bar{\xi}_n| < 3\frac{\sigma}{\sqrt{n}}\right) \approx 0.997,$$

i.e. the inequality

$$|z - \bar{\xi}_n| < 3\frac{\sigma}{\sqrt{n}} \tag{5}$$

is fulfilled with a probability sufficiently close to unity (≈ 0.997).

In practice, if σ is unknown, but in any case finite, then it is estimated for $n > 10$ by the formula

$$\sigma \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi}_n)^2}. \tag{6}$$

Evaluation of Definite Integrals. Consider the integral

$$I = \int_0^1 f(x)\,dx, \tag{7}$$

which may also be an improper integral but such that the integral $\int_0^1 f^2(x)\,dx$ also exists.

Let η be a random quantity distributed uniformly on the interval [0, 1], that is, which has the distribution density

$$p_\eta(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1, \\ 0, & x \notin [0, 1]. \end{cases}$$

Then $\xi = f(\eta)$ will also be some random quantity*, and by the definition of expectation

$$\mathbf{M}[\xi] = \int_0^1 f(x) p_\eta(x)\, dx = \int_0^1 f(x)\, dx = I.$$

Thus,

$$I \approx \bar{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} f(\eta_i), \tag{8}$$

where η_i are independent realizations of the random quantity η. Since, by the supposition, along with (7) the integral $\int_0^1 f^2(x)\, dx$ exists, we have

$$\sigma^2 = \mathbf{D}[f(\eta)] = \mathbf{M}[f^2(\eta)] - (\mathbf{M}[f(\eta)])^2$$

$$= \int_0^1 f^2(x)\, dx - \left(\int_0^1 f(x)\, dx \right)^2 < \infty. \tag{9}$$

The quantity σ is estimated either by the formula (6) or, proceeding from (9), by finding the upper limit of $\int_0^1 f^2(x)\, dx$ and the lower limit of $\left(\int_0^1 f(x)\, dx \right)^2$, say, by zero. The error $I - \bar{\xi}_n$ is estimated by the formula (5) with probability ≈ 0.997.

Multiple integrals are evaluated by the Monte Carlo method in an analogous manner. For instance,

---

* At the points of the interval [0, 1] in the neighbourhood of which the function f is unbounded, we formally set $f = 0$. It is assumed that there is a finite number of such points on [0, 1].

$$\int_0^1 \int_0^1 f(x, y)\, dx\, dy \approx \frac{1}{n} \sum_{i=1}^{n} f(\eta_i, \theta_i), \tag{10}$$

where η_i and θ_i are independent realizations of random quantities η and θ which are distributed uniformly on [0, 1].

In modern computers, random quantities (random numbers) which are distributed uniformly on the interval [0, 1] are specified with the aid of special physical transducers or programs. When using programs, these numbers are called *pseudorandom*. Although practically they possess the statistical characteristics inherent in random numbers, strictly speaking, they are not random.

Comparison with Quadrature Formulas. Consider the following quadrature formula for the double integral:

$$\int_0^1 \int_0^1 f(x, y)\, dx\, dy \approx \sum_{i=0}^{N} \sum_{k=0}^{N} q_{ik} f\left(\frac{i}{N}, \frac{k}{N}\right), \tag{11}$$

where q_{ik} are weights. Let us assume that the error of the formula (11) behaves approximately as $c_1 N^{-m}$ (here and henceforth $c_1, c_2, \ldots$ are certain constants). This behaviour of the error of the quadrature formula is typical. We made sure of this in the one-dimensional case considered in two previous sections.

Let $\varepsilon > 0$ be the assigned accuracy of evaluating the integral. Setting $c_1 N^{-m} \approx \varepsilon$, we find the required $N \approx c_2 \varepsilon^{-1/m}$. Consequently, the number N_f of the values of the function f in the quadrature formula to be computed in ν-dimensional case has the magnitude $N_f \approx N^\nu \approx c_3/\varepsilon^{\nu/m}$.

But according to the Monte Carlo method it must be $3\sigma/\sqrt{n} \approx \varepsilon$, that is, to obtain the same accuracy with probability close to unity, the Monte Carlo method requires the computation of $n \approx c_4/\varepsilon^2$ values of the function f regardless of the dimension ν. For instance, $\nu = 15$ and $m = 3$. We have $N_f \approx c_3/\varepsilon^5$ and $n \approx c_4/\varepsilon^2$.

Thus, when using the Monte Carlo method in a many-dimensional case, the number of the values of the integrand to be computed increases much slower with respect to $1/\varepsilon$ than in quadrature formulas, ε being the accuracy of evaluating the integral. This property expresses the first advantage of the Monte Carlo method.

The second advantage of this method is that its accuracy is independent of the smoothness of the integrand.

The third advantage is that it can be easily adapted to the shape of the region of integration. For instance, we set

$$\iint_{\Omega} f(P)\,dP = \iint_{R} f^*(P)\,dP,$$

where R is the square containing the given domain Ω (Fig. 10),

$$f^*(P) = \begin{cases} f(P), & P \in \Omega, \\ 0, & P \notin \Omega. \end{cases}$$

Then the formula of the form (10) is applied.

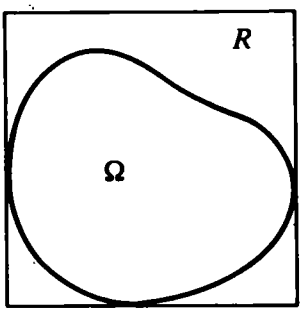


Fig. 10

The shortcoming of the Monte Carlo method is the probabilistic nature of the result, that is, the absence of strict error estimates tending to zero as $n \to \infty$.

Sec. 18. Numerical Methods of Solving Cauchy's Problem for Ordinary Differential Equations

Consider the first-order differential equation

$$u' = f(x, u). \tag{1}$$

In the course of differential equations the following theorem is proved.

Theorem 1. *If the function $f(x, u)$ is continuously differentiable m times with respect to the variables x and u on the two-dimensional closed domain $\bar{G}$, $f \in C_m(\bar{G})$*, then any solution $u(x)$ of the equation* (1) *contained in $\bar{G}$ is $m + 1$ times continuously differentiable with respect to x.*

* See Item 3 of Introduction.

The smoothness of solutions of the equation (1) is verified by Theorem 1.

For the sake of definiteness, let $\bar{G}$ be a closed rectangle:

$$\bar{G} = \{(x, u): x_0 \leqslant x \leqslant x_0 + l, a \leqslant u \leqslant b\}.$$

Cauchy's Problem. Find the solution $u(x)$ of the equation (1) which satisfies the initial condition

$$u(x_0) = u_0, \tag{2}$$

where u_0 is a given number, $a < u_0 < b$.

Suppose that $f \in C_1(\bar{G})$ and there exists the solution $u(x)$ of Cauchy's problem (1), (2) defined on the interval $[x_0, x_0 + l]$, and

$$a < u(x) < b, \ x \in [x_0, x_0 + l]. \tag{3}$$

The uniqueness of the solution of Cauchy's problem follows from the condition $f \in C_1(\bar{G})$. By Theorem 1, $u \in C_2[x_0, x_0 + l]$, and it is obvious that

$$\begin{aligned} u'(x) &= f(x, u(x)), \\ u''(x) &= f'_x(x, u(x)) + f'_u(x, u(x)) f(x, u(x)). \end{aligned} \tag{4}$$

We shall seek an approximate solution of Cauchy's problem (1), (2) on a finite set of the points belonging to the interval $[x_0, x_0 + l]$. This set is called the *grid*. We choose the grid $\omega_h = \{x_j\}_{j=0}^{N}$, where $x_j = x_0 + jh$, $h = l/N$, N is natural.

The function φ defined on the grid ω_h is called the *grid function*. Putting $\varphi_j = \varphi(x_j)$, we introduce the following norm in the linear space of grid functions:

$$\|\varphi\|_h = \max_{0 \leqslant j \leqslant N} |\varphi_j|. \tag{5}$$

The solution $u(x)$ of Cauchy's problem (1), (2) on the interval $[x_0, x_0 + l]$ is defined, in particular, on the grid ω_h (we set $u_j = u(x_j)$). By $\|u\|_h$ we shall imply the norm of the grid function which coincides with $u(x)$ on ω_h.

Let us prove the following auxiliary lemma.

Lemma 1. *Let* $\alpha > 0, \beta \geqslant 0$, *and* $\varepsilon_0 = 0$,

$$|\varepsilon_{j+1}| \leqslant (1 + \alpha)|\varepsilon_j| + \beta, \quad j = 0, 1, \ldots, N - 1. \tag{6}$$

Then for $k = 0, 1, \ldots, N$ *the following inequality holds:*

$$|\varepsilon_k| \leqslant \frac{\beta(e^{k\alpha} - 1)}{\alpha}. \tag{7}$$

Proof. For $k = 0$ the inequality (7) is obvious. Suppose that (7) is true for some $k = m$, $0 \leqslant m < N$. Then, according to (6),

$$|\varepsilon_{m+1}| \leqslant (1 + \alpha)\frac{\beta(e^{m\alpha} - 1)}{\alpha + \beta} = \frac{\beta((1 + \alpha)e^{m\alpha} - 1)}{\alpha} \leqslant \frac{\beta(e^{(m+1)\alpha} - 1)}{\alpha},$$

since $1 + \alpha < e^{\alpha} = 1 + \alpha + \alpha^2/2 + \ldots$. Hence, by induction, there follows (7) for $k = 0, 1, \ldots, N$.

Euler's Method. The approximate solution y of Cauchy's problem (1), (2) is computed on the grid ω_h by the formulas

$$\begin{aligned} y_0 &= u_0, \\ y_{j+1} &= y_j + hf(x_j, y_j) + \eta_j, \quad j = 0, 1, \ldots, \end{aligned} \tag{8}$$

where η_j is a round-off error including the error due to the evaluation of the function f. We shall assume that the approximate solution y is also found in the rectangle $\bar{G}$, that is,

$$a \leqslant y_j \leqslant b, \quad j = 0, 1, \ldots, N. \tag{9}$$

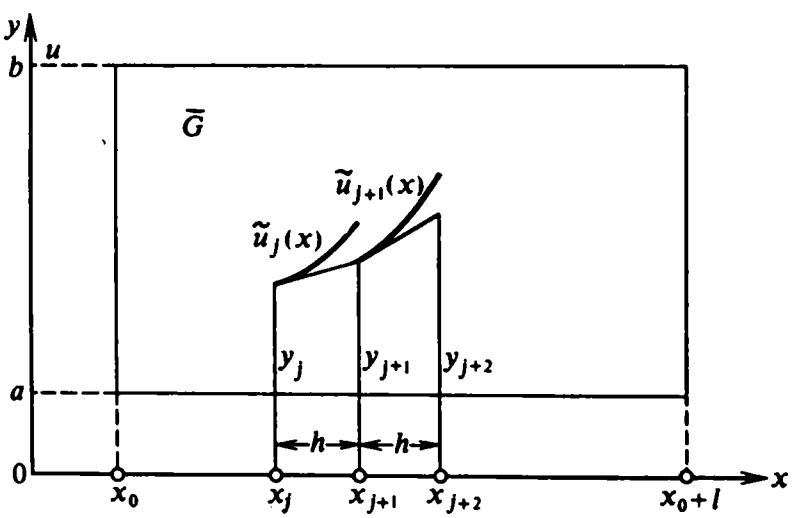


Fig. 11

For $\eta_j = 0$ the passage from y_j to y_{j+1} can be interpreted geometrically as a move along the tangent at the point (x_j, y_j) to the integral curve $\tilde{u}_j(x)$ of the equation (1) which passes through this point over a step h in the direction of the x-axis (Fig. 11). In the absence of the

round-off error the next step will be along the tangent to (generally speaking) another integral curve $\tilde{u}_{j+1}(x)$ passing through the point (x_{j+1}, y_{j+1}) and so on. The initial point with which the construction of an approximate solution begins is the point (x_0, y_0), where $y_0 = u_0$ is the given initial value of the unknown solution $u(x)$ of the equation (1).

Suppose that the round-off error satisfies the condition

$$\max_{0 \leqslant j \leqslant N-1} |\eta_j| \leqslant c_0 h^2, \tag{10}$$

where c_0 is independent of h.

Let us introduce the notation: $\varepsilon_j = u_j - y_j$ (the error of the approximate solution),

$$M_0 = \max_{\bar{G}} |f(x, u)|, \quad M_1 = \max_{\bar{G}} |f'_x(x, u)|,$$
$$M_2 = \max_{\bar{G}} |f'_u(x, u)|. \tag{11}$$

We shall assume that $M_2 > 0$. The case $M_2 = 0$ which means that $f(x, u)$ is independent of u is reduced to evaluating integrals by the methods discussed in Sec. 15.

According to Taylor's formula, for the exact solution $u(x)$ of Cauchy's problem (1), (2), we have

$$u_0 = u_0, \quad u_{j+1} = u_j + hf(x_j, u_j) + \frac{h^2}{2} u''(\xi_j), \tag{12}$$

where $x_j < \xi_j < x_{j+1}$, $j = 0, 1, \ldots, N-1$. Subtracting the equations (12) and (8), and then applying Lagrange's formula of finite increments to the difference $f(x_j, u_j) - f(x_j, y_j)$, we obtain

$$\varepsilon_0 = 0, \quad \varepsilon_{j+1} = \varepsilon_j + hf'_u(x_j, \theta_j)\varepsilon_j + \frac{h^2}{2} u''(\xi_j) - \eta_j,$$

where θ_j lies between u_j and y_j, $j = 0, 1, \ldots, N-1$.

Hence taking into account (4), (10), and (11), we find $\varepsilon_0 = 0$,

$$|\varepsilon_{j+1}| \leqslant (1 + hM_2)|\varepsilon_j| + h^2 M,$$

where $j = 0, 1, \ldots, N-1$,

$$M = c_0 + \frac{M_1 + M_0 M_2}{2}. \tag{13}$$

Consequently, by Lemma 1,

$$\|\varepsilon\|_h \leqslant \frac{h^2 M}{hM_2} (e^{NhM_2} - 1) = h \frac{M}{M_2} (e^{lM_2} - 1). \tag{14}$$

Thus, we have obtained the estimate of the error of the approximate solution y of Cauchy's problem (1), (2) found by Euler's method. This solution is expressed in terms of known quantities.

Two conclusions follow from the aforegoing.

(1) In the absence of round-off errors, the local error of Euler's method, that is, the error over one step h due to the move along the tangent to the integral curve passing through the point (x_j, y_j), but not along the integral curve itself, is the quantity $O(h^2)$. This follows from the expansion of the form (12) by Taylor's formula for solving the equation (1) which passes through the point (x_j, y_j). According to (14), the total error or, more precisely, the maximum error of the solution y on ω_h is equal to $O(h)$, i.e. is by a unity worse in order* than the local error.

(2) The total error is not made worse in order with respect to h if there is a round-off error of the same order as the local error.

The main shortcoming of Euler's method is a low accuracy with respect to h. It is a method of the first order of accuracy.

There are several ways of constructing the numerical methods of solving Cauchy's problem of a higher order of accuracy with respect to h. One of them is based on using the expansion of solution by Taylor's formula (or expansion in a series). But in practice we prefer those methods which only require the right-hand side of the equation (1), and not its derivatives, to be evaluated. Among them are the Runge-Kutta methods and the Adams method.

The Runge-Kutta Methods. We shall consider here three methods.

1. The method of the second order of accuracy (which is also called the **predictor-corrector):**

$$y^*_{j+1} = y_j + hf(x_j, y_j) \quad (y_0 = u_0), \tag{15}$$

$$y_{j+1} = y_j + h\frac{f(x_j, y_j) + f(x_{j+1}, y^*_{j+1})}{2}. \tag{16}$$

The formula (15) predicts the "rough" value of y^*_{j+1} by Euler's method, while the formula (16) corrects the value of the approximate solution at the point x_{j+1}. Let us make sure of it by the following example.

Suppose that $f \in C_2(\bar{G})$. Substituting $x_{j+1} = x_j + h$ into (16) and its expression (15) for y^*_{j+1}, we obtain

$$y_{j+1} = y_j + h\frac{f(x_j, y_j) + f(x_j + h, y_j + hf(x_j, y_j))}{2}.$$

* It is possible to prove that in the general case the indicated order is exact.

Regarding $f(x_j + h, y_j + hf(x_j, y_j))$ as a function only of h (with x_j and y_j being fixed), by Taylor's formula, we find*

$$f(x_j + h, y_j + hf(x_j, y_j)) = f(x_j, y_j) + h(f'_x(x_j, y_j) + f'_u(x_j, y_j) f(x_j, y_j)) + O(h^2),$$

and, consequently,

$$y_{j+1} = y_j + hf(x_j, y_j) + \frac{h^2}{2}(f'_x(x_j, y_j) + f'_u(x_j, y_j) f(x_j, y_j)) + O(h^3). \quad (17)$$

Under the condition that $f \in C_2(\bar{G})$, the solution $v(x)$ of the differential equation (1) which passes through the point (x_j, y_j) is, according to Theorem 1, three times continuously differentiable with respect to x. Therefore taking into consideration the similar expressions (4) for $v'(x)$ and $v''(x)$, by Taylor's formula, we get

$$v_{j+1} = v(x_j) + hv'(x_j) + \frac{h^2}{2} v''(x_j) + O(h^3) = y_j + hf(x_j, y_j) + \frac{h^2}{2}(f'_x(x_j, y_j) + f'_u(x_j, y_j) f(x_j, y_j)) + O(h^3), \quad (18)$$

where $v_{j+1} = v(x_{j+1})$, the quantities $O(h^3)$ in (17) and (18), generally speaking, being not identical. On the basis of (17) and (18), the error due to the drift of y_{j+1} from the integral curve passing through the point (x_j, y_j), i.e. the difference $v_{j+1} - y_{j+1}$, is $O(h^3)$. Thus, the local error of this method has the third order with respect to h, and, consequently, the total error $\|u - y\|_h$ is equal to $O(h^2)$, which is proved in the same way as by Euler's method.

2. The method of the second order of accuracy (also called the **improved Euler's method**):

$$y_{j+1/2} = y_j + \frac{h}{2} f(x_j, y_j) \quad (y_0 = u_0),$$
$$y_{j+1} = y_j + hf\left(x_j + \frac{h}{2}, \; y_{j+1/2}\right). \quad (19)$$

---

* The quantity f'_u denotes the partial derivative with respect to the second argument of the function f which in (1) was denoted by u.

Its local error is investigated just in the same way as in the preceding method.

3. The Runge-Kutta method of the fourth order of accuracy (the most frequently used in practice)

$$y_{j+1} = y_j + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \quad (y_0 = u_0), \tag{20}$$

where

$$k_1 = hf(x_j, y_j), \qquad k_2 = hf\left(x_j + \frac{h}{2}, \; y_j + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x_j + \frac{h}{2}, \; y_j + \frac{k_2}{2}\right), \quad k_4 = hf(x_j + h, y_j + k_3).$$

If $f \in C_4(\bar{G})$, the local error of this method is $O(h^5)$, and the total error, that is, the quantity $\|u - y\|_h$, equals $O(h^4)$.

The Adams Method. Let there be known the approximate values $y_{j-m}, \ldots, y_{j-1}, y_j$ of the solution of Cauchy's problem (1), (2) with step h. Let us designate

$$f_j = f(x_j, y_j). \tag{21}$$

We construct Lagrange's interpolation polynomial of degree m (see Sec. 4):

$$L_m(x) = \sum_{i=0}^{m} p_{mi}(x) f_{j-i} \tag{22}$$

satisfying the conditions

$$L_m(x_{j-i}) = f_{j-i}, \quad i = 0, 1, \ldots, m.$$

Since the exact solution of Cauchy's problem (1), (2) is fulfilled by the equality

$$u_{j+1} = u_j + \int_{x_j}^{x_{j+1}} f(x, u(x))\, dx,$$

where $x_{j+1} = x_j + h$ and $u_k = u(x_k)$, it is natural to set

$$y_{j+1} = y_j + \int_{x_j}^{x_{j+1}} L_m(x)\, dx.$$

Then with (22) in mind, we shall have the following expression for y_{j+1}:

$$y_{j+1} = y_j + h \sum_{i=0}^{m} \alpha_{mi} f_{j-i}, \tag{23}$$

where f_{j-i} are the values of the right-hand member of the equation (1) which are found by the formula (21), and

$$\alpha_{mi} = \frac{1}{h} \int_{x_j}^{x_{j+1}} p_{mi}(x)\, dx, \quad i = 0, 1, \ldots, m,$$

are quite definite constants independent of both h and j. In particular,

$$\alpha_{00} = 1 \quad \text{(Euler's method, } m = 0\text{)},$$

$$\alpha_{10} = \frac{3}{2}, \quad \alpha_{11} = -\frac{1}{2},$$

$$\alpha_{20} = \frac{23}{12}, \quad \alpha_{21} = -\frac{4}{3}, \quad \alpha_{22} = \frac{5}{12},$$

$$\alpha_{30} = \frac{55}{24}, \quad \alpha_{31} = -\frac{59}{24}, \quad \alpha_{32} = \frac{37}{24}, \quad \alpha_{33} = -\frac{3}{8}.$$

It is proved that if $f \in C_{m+1}(\bar{G})$ and there are known $m+1$ values $y_0, y_1, \ldots, y_m$ of the approximate solution at the points $x_0, x_1, \ldots, x_m = x_0 + mh$, and

$$|u_j - y_j| \leqslant ch^{m+1}, \quad j = 0, 1, \ldots, m,$$

where $u_j = u(x_j)$ are the values of the exact solution of the problem (1), (2) and c is independent of h, then the total error of the approximate solution y found at the remaining points of the grid ω_h by the Adams method (23) is equal to $O(h^{m+1})$.

Comparing the Runge-Kutta Methods with the Adams Method. Let us compare the Runge-Kutta methods (20) with the Adams method (23) ($m = 3$) which possess the fourth order of accuracy. When using the method (20), it is required to compute four times the values of the function f for each step h, and, consequently, if the function $f(x, u)$ is sufficiently composite, then this may involve a considerable number of operations on the computer. In the method (23) only one new value of f is computed per step. But the Runge-Kutta method (20) is independent, that is,

we may employ this method having only the initial value y_0, and it is possible to change the step h at any instant. And the Adams method requires a knowledge of the approximate solution at several initial points of the grid and makes it impossible to change the step. In practice these methods are often combined: the initial section is computed by the Runge-Kutta method and then the Adams method is applied. If there is a necessity to change the step h during the process of computation, then we proceed in a similar manner.

Application of Runge's Rule. Suppose that $f \in C_{k+1}(\bar{G})$ and the approximate solution of Cauchy's problem (1), (2) is found by the Runge-Kutta method of the kth order of accuracy (for $k = 1$ by Euler's method). Then it is possible to obtain the following relationship on the grid ω_h:

$$u = y_h + v(x)h^k + r(x, h), \tag{24}$$

where u is the exact solution of the problem (1), (2), y_h is the approximate solution (the subscript h indicates the step with which it is computed by the Runge-Kutta method), $v(x)$ is a function of x independent of h, and $|r(x, h)| \leqslant c'h^{k+1}$, c' is independent of h.

Therefore, on computing the approximate solution three times with steps $2h$, h, and $h/2$, it is possible to estimate the error by Runge's rule at the nodes of the grid ω_{2h} at which the condition (16.13) is fulfilled for the values of the approximate solutions y_{2h}, y_h, and $y_{h/2}$ (corresponding to one and the same x):

$$u - y_{h/2} \approx \frac{y_{h/2} - y_h}{2^k - 1}.$$

Besides, it is possible to find the Richardson improved solution

$$y_h^* = \frac{2^k y_{h/2} - y_h}{2^k - 1}$$

which has the error $O(h^{k+1})$ instead of $O(h^k)$.

Remark 1. The relationship (24) is preserved if the approximate solution is found by the formulas (15), (16), or (19), or (20) not ideally but with round-off errors (the same as in Euler's method (8)) which on each step are bounded by the quantity ch^{k+2}, where k is the order of the total accuracy of the method , and c is independent of h. It is difficult to check directly and maintain this condition for the round-off error on a computer. It is more natural to check the fulfilment of the condition (16.13) at the final or some intermediate point of integration, say, at the point

$x_0 + l$ or $x_0 + l/2$. If the round-off errors begin prevailing over the errors of the method itself, then the condition (16.13) is roughly violated. Then we have either to increase the step h or to change the calculation program so that a double or triple precision is used. But this is not always possible.

Remark 2. In the case of Cauchy's problem for the normal system of differential equations

$$\mathbf{u}' = \mathbf{f}(x, \mathbf{u}), \mathbf{u}(x_0) = \mathbf{u}_0$$

the above-discussed Euler, Runge-Kutta, and Adams approximate integration methods formally remain the same, the functions u, y, f, and the quantities k_i in the formula (20) being replaced by the vector functions $\mathbf{u}$, $\mathbf{y}$, $\mathbf{f}$, and vectors $\mathbf{k}_i$, respectively. The above requirements concerning the smoothness of the function f remain the same for each component of the vector function $\mathbf{f}$ and the guaranteed accuracy of the order with respect to h will be the same (as in the case of one equation (1)) for the components of the approximate solution. Runge's rule is applied to each component separately.

On the Existence of Solution of an Original Problem. Let us consider the Cauchy problem

$$u' = u^2, \quad u(0) = 1. \tag{25}$$

If for solving this problem we use, say, Euler's or Runge-Kutta method (20) with any step $h > 0$, then an approximate solution can be, in principle, computed for any number of steps, that is, for arbitrarily large values of the argument x.

But the only exact solution of the problem (25) is the function $u(x) = (1 - x)^{-1}$ defined for nonnegative x's only on the interval $[0, 1)$.

Thus, the existence of an approximate solution of Cauchy's problem on some interval does not yet guarantee the existence of the solution of the original differential problem on this interval.

Let us give two theorems (without proof) which enable us to decide whether there is a solution to Cauchy's problem (1), (2) in a given rectangle $\bar{G}$ in terms of the approximate solution obtained by Euler's method. Let

$$\bar{c} = \frac{M}{M_2}(e^{lM_2} - 1) + M_0,$$

where M_0, M_2, and M are defined by the formulas (11) and (13).

Theorem 2. *If $f \in C_1(\bar{G})$, and the approximate solution which is found by the formulas* (8) *for certain fixed $h = l/N$ obey the inequalities*

$$a + \bar{c}h < \max_{0 \leqslant j \leqslant N} y_i \leqslant \max_{0 \leqslant j \leqslant N} y_j < b - \bar{c}h, \tag{26}$$

then the solution $u(x)$ of Cauchy's problem (1), (2) *exists on the interval $[x_0, x_0 + l]$ and satisfies the conditions* (3).

Theorem 3. *If $f \in C_1(\bar{G})$, and on the interval $[x_0, x_0 + l]$ there exists the solution $u(x)$ of Cauchy's problem* (1), (2) *which satisfies the conditions* (3), *then there is $h_0 > 0$ such that for any $h = l/N < h_0$ the inequalities* (26) *for the approximate solution which is found by the formulas* (8) *will be automatically fulfilled.*

According to Theorem 2, for the solution of Cauchy's problem (1), (2) which satisfies the conditions (3) to exist on the interval $[x_0, x_0 + l]$, it is sufficient that the approximate solution found by Euler's method satisfy the inequalities (26) only for one arbitrary value of h. Theorem 3 guarantees that if the indicated solution of Cauchy's problem (1), (2) exists, then the inequalities (26) will be fulfilled for all sufficiently small $h = l/N$.

Chapter 3

NUMERICAL METHODS OF LINEAR ALGEBRA

The numerical methods of linear algebra include the numerical methods of solving systems of linear algebraic equations, matrix inversion, computing determinants, and finding the eigenvalues and eigenvectors of matrices.

Methods of solving systems of linear algebraic equations are subdivided into two groups. The first group comprises so-called exact or direct methods, that is, algorithms enabling us to obtain the solution of a system after a finite number of arithmetic operations. Among them are: Cramer's rule for finding the solution of a system with the aid of determinants, the Gauss elimination method, and the sweep method. The second group includes approximate methods, the iteration methods of solving systems of linear algebraic equations, in particular. Cramer's rule is not used with computers, since it requires a considerably greater number of arithmetic operations than the Gauss method. The Gauss elimination method is usually used with computers when solving systems up to the order of 10^3, and iteration methods are used up to the order of 10^6. The sweep method is employed for solving the important class of specific systems of linear equations with a tridiagonal matrix frequently occurring in various applications.

The Gauss elimination method of solving systems of linear equations, matrix inversion, and computing determinants is set forth in Sec. 19. Section 20 introduces the notion of the norm of a matrix used in Sec. 21 when considering the method of simple iterations. One of the important characteristics of the matrix of a system of linear algebraic equations is its measure of the sensitivity of the solution of a system to a perturbation of the right-hand side. An example of an ill-conditioned matrix is given in Sec. 20, and a well-conditioned matrix is presented in Sec. 21. Section 22 is dedicated to the sweep method. Some problems of eigenvalues, mainly for symmetric matrices, are considered in Sec. 23.

Sec. 19. The Gauss Elimination Method

For the sake of simplicity, let us consider the fourth-order system of linear algebraic equations

$$\begin{aligned} a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + a_{13}^{(0)}x_3 + a_{14}^{(0)}x_4 &= a_{15}^{(0)}, \\ a_{21}^{(0)}x_1 + a_{22}^{(0)}x_2 + a_{23}^{(0)}x_3 + a_{24}^{(0)}x_4 &= a_{25}^{(0)}, \\ a_{31}^{(0)}x_1 + a_{32}^{(0)}x_2 + a_{33}^{(0)}x_3 + a_{34}^{(0)}x_4 &= a_{35}^{(0)}, \\ a_{41}^{(0)}x_1 + a_{42}^{(0)}x_2 + a_{43}^{(0)}x_3 + a_{44}^{(0)}x_4 &= a_{45}^{(0)}. \end{aligned} \tag{1}$$

Suppose that the coefficient $a_{11}^{(0)}$, called the *leading element* of the first row, is not equal to zero. Dividing the first equation of (1) by $a_{11}^{(0)}$, we get a new equation

$$x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 = a_{15}^{(1)}, \tag{2}$$

where $a_{1j}^{(1)} = a_{1j}^{(0)}/a_{11}^{(0)}$, $j = 2, 3, 4, 5$.

Let us estimate the unknown x_1 from each equation of the system (1), beginning with the second, by subtracting the equation (2) multiplied by the coefficient of x_1 of the appropriate equation. The transformed equations have the form

$$\begin{aligned} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 &= a_{25}^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 &= a_{35}^{(1)}, \\ a_{42}^{(1)}x_2 + a_{43}^{(1)}x_3 + a_{44}^{(1)}x_4 &= a_{45}^{(1)}, \end{aligned} \tag{3}$$

where $a_{ij}^{(1)} = a_{ij}^{(0)} - a_{1j}^{(1)}a_{i1}^{(0)}$, $i = 2, 3, 4$, $j = 2, 3, 4, 5$.

Let us assume now that the leading element of the second row, that is, the coefficient $a_{22}^{(1)}$, is also different from zero. Then dividing the first equation of (3) by $a_{22}^{(1)}$, we obtain the equation

$$x_2 + a_{23}^{(2)}x_3 + a_{24}^{(2)}x_4 = a_{25}^{(2)}, \tag{4}$$

where $a_{2j}^{(2)} = a_{2j}^{(1)}/a_{22}^{(1)}$, $j = 3, 4, 5$.

Eliminating the unknown x_2 from the last two equations in (3) in terms of the equation (4), we arrive at the equations

$$\begin{aligned} a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 &= a_{35}^{(2)}, \\ a_{43}^{(2)}x_3 + a_{44}^{(2)}x_4 &= a_{45}^{(2)}, \end{aligned} \tag{5}$$

where $a_{ij}^{(2)} = a_{ij}^{(1)} - a_{2j}^{(2)}a_{i2}^{(1)}$, $i = 3, 4$, $j = 3, 4, 5$.

If the leading element of the third row $a_{33}^{(2)}$ is not equal to zero, then dividing the first of the equations (5) by $a_{33}^{(2)}$ and subtracting the obtained

equation multiplied by $a_{43}^{(2)}$ from the second equation, we get

$$x_3 + a_{34}^{(3)}x_4 = a_{35}^{(3)}, \tag{6}$$

$$a_{44}^{(3)}x_4 = a_{45}^{(3)}, \tag{7}$$

where $a_{3j}^{(3)} = a_{3j}^{(2)}/a_{33}^{(2)}$, $a_{4j}^{(3)} = a_{4j}^{(2)} - a_{3j}^{(3)}a_{43}^{(2)}$, $j = 4, 5$.

Finally, if $a_{44}^{(3)} \neq 0$, then dividing the equation (7) by this element, we bring this equation to the form

$$x_4 = a_{45}^{(4)}, \tag{8}$$

where $a_{45}^{(4)} = a_{45}^{(3)}/a_{44}^{(3)}$.

Thus, if the leading elements $a_{11}^{(0)}$, $a_{22}^{(1)}$, $a_{33}^{(2)}$, and $a_{44}^{(3)}$ are different from zero, then the system (1) is equivalent to the following system with a triangular matrix:

$$\begin{aligned} x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 &= a_{15}^{(1)}, \\ x_2 + a_{23}^{(2)}x_3 + a_{24}^{(2)}x_4 &= a_{25}^{(2)}, \\ x_3 + a_{34}^{(3)}x_4 &= a_{35}^{(3)}, \\ x_4 &= a_{45}^{(4)}, \end{aligned} \tag{9}$$

which is obtained by combining the equations (2), (4), (6), and (8). From the system (9) the unknowns x_1, x_2, x_3, and x_4 are found explicitly in the reverse order by the formulas

$$\begin{aligned} x_4 &= a_{45}^{(4)}, \\ x_3 &= a_{35}^{(3)} - a_{34}^{(3)}x_4, \\ x_2 &= a_{25}^{(2)} - a_{23}^{(2)}x_3 - a_{24}^{(2)}x_4, \\ x_1 &= a_{15}^{(1)} - a_{12}^{(1)}x_2 - a_{13}^{(1)}x_3 - a_{14}^{(1)}x_4. \end{aligned} \tag{10}$$

The process of reducing the original system (1) to the triangular form (9) is ordinarily called the *forward procedure,* and finding the unknowns by the formulas (10) is called the *backward procedure* of the Gauss method.

The solution of the system (1) by the Gauss method without using a computer is usually presented in tabular form (see Table 1). The last column in the table is given with the purpose of checking the computations, which is done in the following way. We assign $a_{i6}^{(0)}$ to be equal to the sum of the remaining elements of the ith row taken with opposite sign, that is,

$$a_{i6}^{(0)} = -(a_{i1}^{(0)} + a_{i2}^{(0)} + \ldots + a_{i5}^{(0)}), \quad i = 1, 2, 3, 4.$$

Table 1

$a_{11}^{(0)}$	$a_{12}^{(0)}$	$a_{13}^{(0)}$	$a_{14}^{(0)}$	$a_{15}^{(0)}$	$a_{16}^{(0)}$
$a_{21}^{(0)}$	$a_{22}^{(0)}$	$a_{23}^{(0)}$	$a_{24}^{(0)}$	$a_{25}^{(0)}$	$a_{26}^{(0)}$
$a_{31}^{(0)}$	$a_{32}^{(0)}$	$a_{33}^{(0)}$	$a_{34}^{(0)}$	$a_{35}^{(0)}$	$a_{36}^{(0)}$
$a_{41}^{(0)}$	$a_{42}^{(0)}$	$a_{43}^{(0)}$	$a_{44}^{(0)}$	$a_{45}^{(0)}$	$a_{46}^{(0)}$
1	$a_{12}^{(1)}$	$a_{13}^{(1)}$	$a_{14}^{(1)}$	$a_{15}^{(1)}$	$a_{16}^{(1)}$
	$a_{22}^{(1)}$	$a_{23}^{(1)}$	$a_{24}^{(1)}$	$a_{25}^{(1)}$	$a_{26}^{(1)}$
	$a_{32}^{(1)}$	$a_{33}^{(1)}$	$a_{34}^{(1)}$	$a_{35}^{(1)}$	$a_{36}^{(1)}$
	$a_{42}^{(1)}$	$a_{43}^{(1)}$	$a_{44}^{(1)}$	$a_{45}^{(1)}$	$a_{46}^{(1)}$
	1	$a_{23}^{(2)}$	$a_{24}^{(2)}$	$a_{25}^{(2)}$	$a_{26}^{(2)}$
		$a_{33}^{(2)}$	$a_{34}^{(2)}$	$a_{35}^{(2)}$	$a_{36}^{(2)}$
		$a_{43}^{(2)}$	$a_{44}^{(2)}$	$a_{45}^{(2)}$	$a_{46}^{(2)}$
		1	$a_{34}^{(3)}$	$a_{35}^{(3)}$	$a_{36}^{(3)}$
			$a_{44}^{(3)}$	$a_{45}^{(3)}$	$a_{46}^{(3)}$
			1	x_4	$\bar{x}_4$
		1		x_3	$\bar{x}_3$
	1			x_2	$\bar{x}_2$
1				x_1	$\bar{x}_1$

Then the sum of all the elements of each of the four expanded original rows will be equal to zero. Further, in the process of eliminating the unknowns we perform the same operations on the elements of the last column as on the elements of the appropriate rows, namely, we set

$$a_{kj}^{(k)} = \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)} a_{kj}^{(k)}, \tag{11}$$

where $k + 1 \leqslant j \leqslant 6$, $k + 1 \leqslant i \leqslant 4$, $k = 1, 2, 3, 4$.

In this case, the sum of the elements of the newly obtained expanded rows must remain equal to zero, since these elements are found in terms of linear operations on the previous rows. For example,

$$1 + a_{23}^{(2)} + a_{24}^{(2)} + a_{25}^{(2)} + a_{26}^{(2)} = 0. \tag{12}$$

If computations are carried out with rounding-offs, then the left-hand member of the equality (12) is allowed to deviate from zero within the bounds of round-off errors. A large deviation testifies to the fact that there is a gross error in the calculations.

The unknowns x_4, x_3, x_2, and x_1 are found by the formulas (10), and the quantities $\bar{x}_4$, $\bar{x}_3$, $\bar{x}_2$, and $\bar{x}_1$ by the same formulas in which the numbers $a_{k5}^{(k)}$ are replaced by $a_{k6}^{(k)}$, $k = 4, 3, 2, 1$, and x_i by $\bar{x}_i$. In the absence of round-off errors the sum of the elements in the last four rows of Table 1 must also be equal to zero, that is, $1 + x_i + \bar{x}_i = 0$, $i = 4, 3, 2, 1$. In Table 1 the elements $a_{45}^{(4)}$ and $a_{46}^{(4)}$ are omitted, since they coincide with x_4 and $\bar{x}_4$, respectively.

Example. It is required to solve the system of linear algebraic equations

$$\begin{aligned} x_1 + 0.17x_2 - 0.25x_3 + 0.54x_4 &= 0.3, \\ 0.47x_1 + x_2 + 0.67x_3 - 0.32x_4 &= 0.5, \\ -0.11x_1 + 0.35x_2 + x_3 - 0.74x_4 &= 0.7, \\ 0.55x_1 + 0.43x_2 + 0.36x_3 + x_4 &= 0.9. \end{aligned} \tag{13}$$

The solution is given in Table 2. The computations are carried out with two spare decimal places as compared with the number of decimal places in the coefficients of the given system. The sum of the elements of each row of Table 2, including the element contained in the check column, differs from zero by not more than two unities of the junior decimal place, which is quite permissible. The found solution of the system (13) which is rounded off to two decimal places is: $x_1 = 0.44$, $x_2 = -0.36$, $x_3 = 1.17$, and $x_4 = 0.39$.

Analogously, the Gauss method is applied to solving a system of linear algebraic equations of any order n.

Let there be given a system

$$\sum_{j=1}^{n} a_{ij}^{(0)} x_j = a_{i,\,n+1}^{(0)}, \quad i = 1, 2, \ldots, n. \tag{1*}$$

Table 2

1	0.17	−0.25	0.54	0.3	−1.76
0.47	1	0.67	−0.32	0.5	−2.32
−0.11	0.35	1	−0.74	0.7	−1.20
0.55	0.43	0.36	1	0.9	−3.24
1	0.17	−0.25	0.54	0.3	−1.76
	0.9201	0.7875	−0.5738	0.3590	−1.4928
	0.3687	0.9725	−0.6806	0.7330	−1.3936
	0.3365	0.4975	0.7030	0.7350	−2.2720
	1	0.8559	−0.6236	0.3902	−1.6224
		0.6569	−0.4507	0.5891	−0.7954
		0.2095	0.9128	0.6037	−1.7261
		1	−0.6861	0.8968	−1.2108
			1.0565	0.4158	−1.4724
			1	0.3936	−1.3937
		1		1.1668	−2.1670
	1			−0.3630	−0.6368
1				0.4409	−1.4409

If $a_{11}^{(0)} \neq 0$, and the leading elements $a_{ii}^{(i-1)}$, $i = 2, 3, \ldots, n$, of the rest of the rows obtained in the process of computations are different from zero, then the system (1*) is reduced to the triangular form

$$x_i + \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j = a_{i,n+1}^{(i)}, \quad i = 1, 2, \ldots, n. \tag{9*}$$

The leading elements $a_{ii}^{(i-1)}$ and the coefficients of the system (9*) are found in terms of the formulas (11), where $k + 1 \leqslant j \leqslant n + 1, k + 1 \leqslant i \leqslant n + 1, k = 1, 2, \ldots, n$.

The backward procedure in which the unknowns are determined in the reverse order is realized by the formulas

$$x_n = a^{(n)}_{n,\,n+1},$$

$$x_i = a^{(i)}_{i,\,n+1} - \sum_{j=i+1}^{n} a^{(i)}_{ij} x_j, \quad i = n-1, \quad n-2, \ldots, 1. \qquad (10^*)$$

Counting the Number of Arithmetic Operations Performed. Let us first prove an auxiliary lemma.

Lemma 1. *The following equalities are fulfilled:*

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \qquad (14)$$

Proof. The first equality which expresses the sum of an arithmetic progression is known. The second equality is obvious for $n = 1$. Suppose that it is true for some natural $n = m$. Then

$$\sum_{k=1}^{m+1} k^2 = \sum_{k=1}^{m} k^2 + (m+1)^2$$

$$= \frac{m(m+1)(2m+1)}{6} + (m+1)^2$$

$$= \frac{(m+1)[(m+1)+1][2(m+1)+1]}{6}.$$

Consequently, by induction, the second equality of (14) holds for any natural n.

Let us count the number of arithmetic operations to be performed when solving a system of order n neglecting the checking operations. When eliminating the unknown x_1, the following operations are performed:

(a) a total of n divisions of the second and subsequent coefficients of the first equation and its right-hand member by the coefficient of x_1;

(b) n multiplications and n subtractions when eliminating x_1 from the second and each of the subsequent equations whose number is equal to $n - 1$.

Thus, the total number of arithmetic operations required for eliminating x_1 is equal to

$$Q_1 = n + 2n(n-1) = 2n^2 - n. \qquad (15)$$

When eliminating x_2, we perform the computations similar to those carried out for the elimination of x_1, the only difference being that the operations are performed with the rows of a matrix whose dimension in each direction is less by unity. Therefore when eliminating x_2, the number Q_2 of the arithmetic operations performed is obtained from the expression (15) by substituting $n - 1$ for n:

$$Q_2 = 2(n - 1)^2 - (n - 1) = 2(n - 2 + 1)^2 - (n - 2 + 1).$$

In general, when eliminating the unknown x_i, we perform Q_i arithmetic operations:

$$Q_i = 2(n - i + 1)^2 - (n - i + 1).$$

The total number of arithmetic operations to be performed in the forward procedure amounts to

$$Q_* = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} [2(n - i + 1)^2 - (n - i + 1)].$$

Hence, setting $k = n - i + 1$ and relying on Lemma 1, we find the explicit expression for Q_*:

$$Q_* = \sum_{k=n}^{1} (2k^2 - k) = \sum_{k=1}^{n} (2k^2 - k)$$

$$= \frac{n(n + 1)(2n + 1)}{3} - \frac{n(n + 1)}{2} = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{n}{6}. \quad (16)$$

In the backward procedure (given by the formulas (10) for $n = 4$) the number of multiplications and subtractions taken together is equal to the double number of the elements of the square matrix of order n which lie under the principal diagonal. Simple calculations show that this number is equal to

$$Q^* = n(n - 1) = n^2 - n. \quad (17)$$

Thus, the number of arithmetic operations to be performed when solving a system of linear algebraic equations of order n by the Gauss method in the general case is equal to

$$Q = Q_* + Q^* = \frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n \approx \frac{2}{3}n^3. \quad (18)$$

Then we have to know separately the number of arithmetic operations to be performed on the right-hand members of the equations in the forward procedure. When eliminating x_1, the right-hand side of the first equation is divided by the coefficient of x_1, and when finding the right-hand sides of the remaining equations, one multiplication and one subtraction are needed. Thus, when eliminating x_1, a total of q_1 arithmetic operations is performed on the right-hand sides:

$$q_1 = 1 + 2(n - 1) = 2n - 1.$$

Accordingly, when eliminating x_i, we perform q_i arithmetic operations on the right-hand sides of the equations:

$$q_i = 2(n - i + 1) - 1.$$

Consequently, in the forward procedure the number of arithmetic operations performed on the right-hand members of the equations amounts to

$$q_* = \sum_{i=1}^{n} [2(n - i + 1) - 1] = \sum_{k=1}^{n} (2k - 1)$$

$$= n(n + 1) - n = n^2. \quad (19)$$

Computing Determinants. In the forward procedure of the Gauss method, the matrix $(a_{ij}^{(0)})_{i,j=1}^{n}$ of the given system of linear algebraic equations (1*) is reduced to triangular form if the leading elements of all the rows are different from zero. This is achieved by dividing the rows by their leading elements and by subtracting the rows which are normed in such a way and multiplied by certain numbers from the subsequent rows (this procedure is described in detail for $n = 4$). When dividing some row of the matrix by the leading element, the determinant of the matrix is also divided by this element, and when subtracting some row multiplied by any number from another row of the matrix, the determinant, as is known, remains the same. The determinant of a triangular matrix which is found in the forward procedure and whose elements in the principal diagonal are all unity, that is, the determinant of the system (9*) which coincides with the system (9) for $n = 4$, is equal to unity. It is obtained as the result of dividing the determinant of the matrix $(a_{ij}^{(0)})_{i,j=1}^{n}$, denoted $\det(a_{ij}^{(0)})$, by the leading elements $a_{11}^{(0)}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$ of all the rows. Consequently, if $a_{11}^{(0)} \neq 0, a_{22}^{(1)} \neq 0, \ldots, a_{nn}^{(n-1)} \neq 0$, then

$$\det(a_{ij}^{(0)}) = a_{11}^{(0)} a_{22}^{(1)} \ldots a_{nn}^{(n-1)}. \quad (20)$$

Thus, when solving a system of linear algebraic equations by the Gauss method, we can incidentally compute its determinant. If it is required to find only the determinant of the given matrix $(a_{ij}^{(0)})_{i,j=1}^{n}$, then the computations are carried out according to the scheme corresponding to the forward procedure of the Gauss method, the only difference being that there are no operations on the column of the right-hand sides. According to (16), (19), and (20), the number of arithmetic operations performed in computing $\det(a_{ij}^{(0)})$ by this method is equal to

$$Q_0 = Q_* - q_* + (n-1) = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{n}{6} - n^2 + n - 1$$

$$= \frac{2}{3}n^3 - \frac{n^2}{2} + \frac{5}{6}n - 1 \approx \frac{2}{3}n^3, \quad (21)$$

where Q_* is the number of operations performed in the forward procedure when solving a system of order n, q_* is the number of operations performed in this case on the right-hand sides of the equations, and $n-1$ is the number of multiplications in the formula (20).

Remark 1. If in the course of the forward procedure the leading element of some row turns out to be equal to zero, then the simplest variant of the Gauss method discussed above is unfit for solving the appropriate system or computing the determinant. Below, we consider a more universal variant of the Gauss method with the leading element chosen.

Solving Several Systems with Different Right-hand Members. If there are given several systems of linear algebraic equations of nth order which differ only in their right-hand members, then it is advisable to realize the forward procedure simultaneously with all the columns of the right-hand members (in the same way as with the column of the right-hand members and the additional check column in the foregoing example). The backward procedure for finding the solution of each system is carried out successively. According to (17)-(19), the number of arithmetic operations performed when solving p indicated systems is equal to

$$Q(p) = Q + (p-1)(q_* + Q^*)$$

$$= \frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n + (p-1)(n^2 + n^2 - n)$$

$$= \frac{2}{3}n^3 - \frac{n^2}{2} - \frac{n}{6} + pn(2n-1) \approx \frac{2}{3}n^3 + 2pn^2, \quad (22)$$

where Q is the number of operations required for a complete solution of one system, q_* is the number of operations in the forward procedure on one column of the right-hand members, and Q^* is the number of operations in the backward procedure when finding one solution.

Computing the Inverse of a Matrix. Let there be given a nonsingular matrix $A = (a_{ij}^{(0)})_{i,j=1}^{n}$. From linear algebra it is known that the jth column of the inverse (or reciprocal) matrix A^{-1} coincides with the column $(x_{1j}, x_{2j}, \ldots, x_{nj})^*$, where the asterisk is the transposition sign, and $(x_{1j}, x_{2j}, \ldots, x_{nj})$ is the solution of the system*

$$Ax_j = \delta_j, \tag{23}$$

$\delta_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj})$, δ_{ij} is the Kronecker delta. Therefore, to find the inverse matrix A^{-1}, it suffices to solve n systems of the form (23) for $j = 1, 2, \ldots, n$ which differ only in their right-hand sides. In this case,

$$Q(n) \approx \frac{8}{3} n^3 \tag{24}$$

arithmetic operations are required.

A Scheme with the Main Element Chosen. The foregoing simplest variant of the Gauss method, called the *scheme of unique division,* has the following disadvantages. If the leading element of some row, say, the coefficient $a_{11}^{(0)}$ of x_1, turns out to be equal to zero just in the first equation, then this scheme is formally unfit, although the given system of equations may have a unique solution. Besides, if the determinant of the system is different from zero, but in the process of computation some leading elements are encountered which are sufficiently small when compared with the other elements of the appropriate rows, then as a result of this circumstance the negative effect of round-off errors on the accuracy of the solution increases.

Let us set forth the scheme, called the *scheme with the main element chosen,* which, in case when the determinant of a system is different from zero and there are no round-off errors, always leads to a unique solution and is less sensitive to round-off errors. This scheme differs but little from the scheme described at the beginning of this section.

Let, as before, there be given the system (1). We first fulfill the condition

$$|a_{11}^{(0)}| \geqslant |a_{ij}^{(0)}|, \quad i, j = 1, 2, 3, 4,$$

by interchanging, if necessary, two equations of the system (1), as well as two columns of unknowns with their coefficients, and by appropriate

---

* Here and in the subsequent sections, we shall stick to the notation adopted in the book by Ya. S. Bugrov and S. M. Nikolsky *Fundamentals of Linear Algebra and Analytical Geometry* (in English—Mir Publishers, Moscow, 1982), where the vectors of unknowns, solutions, and right-hand sides are written in line.

renumbering of the coefficients and unknowns. The maximum-modulus coefficient thus obtained denoted by $a_{11}^{(0)}$, when renumbering, is called the *first principal element.*

Then dividing the first equation by $a_{11}^{(0)}$, we reduce it to the form (2). Eliminating x_1 from the second, third, and fourth equations, we get the equations (3).

Then we deal with the system (3) in the same way as with the entire system (1). Namely, if necessary, rearranging two equations and, possibly, interchanging two columns of the unknowns with their coefficients, and carrying out appropriate renumbering, we fulfill the inequalities

$$|a_{22}^{(1)}| \geqslant |a_{ij}^{(1)}|, \quad i, j = 2, 3, 4.$$

In this case, if some columns of unknowns were interchanged, a corresponding rearrangement is carried out in the equation (2). The maximum-modulus coefficient thus obtained, denoted by $a_{22}^{(1)}$, is called the *second principal element.*

Dividing the equation which occupies now the first place in (3) by $a_{22}^{(1)}$, we obtain the equation (4), and eliminating x_2 from the second and third equations, we arrive at the equations (5). If when finding the principal element, the columns are rearranged in the system (5), then appropriate rearrangement and renumbering should be realized in the already separated equations (2) and (4) which enter into the final system (9), and so forth.

If the determinant of the system (1) is different from zero, then the forward procedure will result in a system of the form (9). The backward procedure is carried out by the formulas (10). Returning to the original numbering of the unknowns thus found, we obtain the solution of the given system (1).

Analogously, the Gauss method with the principal element chosen is applied to solving a system of any order n.

Remark 2. If the matrix of the given system is singular, then prior to eliminating some unknown, the principal element by which the first of the remaining equations must be divided will turn out to be equal to zero. This will reveal that the determinant of the given system equals zero. The problem of solving a system of equations with a singular matrix is considered in the course of linear algebra. We shall not dwell on this case in more detail.

Remark 3. The Gauss method with the principal element chosen enables us to compute the determinant (if the latter is not equal to zero) by the formula (20) the right-hand side of which contains the product of the principal elements with the multiplier $(-1)^{\nu}$ added, where ν is the sum

of the numbers of the rows and columns rearranged in the forward procedure at all the steps of reducing the matrix to triangular form. And if the determinant is equal to zero, then this fact will be revealed in the course of computations, since a principal element equal to zero will occur at some step.

Remark 4. In practice, when reducing a system to triangular form, operations are performed only with the elements of its matrix, and the values of the unknowns thus obtained are written out in the backward procedure as, for instance, it has been illustrated by Tables 1 and 2.

Remark 5. There is the following variant of the scheme with the principal element chosen. The recurrent (say, the first) principal element thus found is left at its place (it is not brought to the upper left corner of the matrix by rearranging its rows and columns). The unknown corresponding to the chosen principal element is eliminated from all the equations among the coefficients of which the principal elements have not yet been isolated. The next principal (maximum by modulus) element is found among the coefficients of the equations from which the next unknown is eliminated, and so on. If the determinant of the given system of order n is not equal to zero, then over n steps an equivalent system will be obtained which can be reduced to triangular form by rearranging its equations and renumbering the unknowns.

Sec. 20. Norms and Condition of Matrices

In the linear space $\mathbf{R}^n$ of n-dimensional vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ let us introduce the norm, which can be done in two ways:

$$\|\mathbf{x}\| = \begin{cases} \|\mathbf{x}\|_1 = \max\limits_{1 \leqslant j \leqslant n} |x_j| \quad \text{or} \\ \\ \|\mathbf{x}\|_2 = (\mathbf{x}, \mathbf{x})^{1/2} = \left(\sum\limits_{j=1}^{n} x_j^2 \right)^{1/2}. \end{cases} \tag{1}$$

Let us verify the fulfilment of the axioms of norm (see Item 7 of Introduction) for the norm $\|\mathbf{x}\|_1$.

For any $\mathbf{x} \in \mathbf{R}^n$ we have $\max\limits_{1 \leqslant j \leqslant n} |x_j| \geqslant 0$, that is, $\|\mathbf{x}\|_1 \geqslant 0$. In this case, $\max\limits_{1 \leqslant j \leqslant n} |x_j| = 0$ or $\|\mathbf{x}\|_1 = 0$ if and only if $\mathbf{x} = 0$. Besides, $\max\limits_{1 \leqslant j \leqslant n} |\alpha x_j| = \max\limits_{1 \leqslant j \leqslant n} |\alpha| \, |x_j| = |\alpha| \max\limits_{1 \leqslant j \leqslant n} |x_j|$, that is, $\|\alpha \mathbf{x}\|_1 = |\alpha| \, \|\mathbf{x}\|_1$, where α is any number. Finally, let there be given two arbitrary vectors $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$. Then there is a number k such that

$\max\limits_{1 \leqslant j \leqslant n} |x_j + y_j| = |x_k + y_k|$. Consequently, $\max\limits_{1 \leqslant j \leqslant n} |x_j + y_j| = |x_k + y_k| \leqslant |x_k| + |y_k| \leqslant \max\limits_{1 \leqslant j \leqslant n} |x_j| + \max\limits_{1 \leqslant j \leqslant n} |y_j|$, i.e. $\|\mathbf{x} + \mathbf{y}\|_1 \leqslant \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1$.

The axioms for the norm $\|\mathbf{x}\|_2$ introduced in terms of a scalar product are verified in the course of linear algebra.

Obviously, the first and second norms of any $\mathbf{x} \in \mathbf{R}^n$ satisfy the inequalities

$$\|\mathbf{x}\|_1 \leqslant \|\mathbf{x}\|_2 \leqslant \sqrt{n}\|\mathbf{x}\|_1. \tag{2}$$

In the linear space of square number matrices of order n

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

let us specify the norm in the following way:

$$\|A\| = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}, \tag{3}$$

where $\|\mathbf{x}\|$ and $\|A\mathbf{x}\|$ are the above-introduced norms of the vectors $\mathbf{x}$ and $A\mathbf{x}$, respectively. The norm of the matrix specified by the formula (3) is said to be *consistent* with the norm of a vector.

In particular, if $A = E$, that is, if A is a unit matrix, then (since $E\mathbf{x} = \mathbf{x}$ for any $\mathbf{x} \in \mathbf{R}^n$ and, consequently, $\|A\mathbf{x}\| = \|E\mathbf{x}\| = \|\mathbf{x}\|$) it follows from (3) that

$$\|E\| = 1. \tag{4}$$

We shall denote the norm of matrix A by $\|A\|_m$ in the case when the norm $\|\mathbf{x}\| = \|\mathbf{x}\|_m$, $m = 1, 2$, is taken for the vector.

It is possible to prove that

$$\|A\|_1 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|, \tag{5}$$

$$\|A\|_2 \leqslant \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{1/2}, \tag{6}$$

and if matrix A is symmetric, i.e. $A = A^*$ (where A^* is the matrix transposed to A), then

$$\|A\|_2 = \max_{1 \leqslant i \leqslant n} |\lambda_i(A)|, \quad \|A^{-1}\|_2 = \frac{1}{\min\limits_{1 \leqslant i \leqslant n} |\lambda_i(A)|}, \tag{7}$$

where $\lambda_i(A)$ are the eigenvalues of matrix A.

Let A and B be square number matrices of order n, and let $A + B$ be their sum, $\mathbf{x} \in \mathbf{R}^n$. Since $\|(A + B)\mathbf{x}\| = \|A\mathbf{x} + B\mathbf{x}\| \leqslant \|A\mathbf{x}\| + \|B\mathbf{x}\|$, we have

$$\|A + B\| = \sup_{\mathbf{x} \neq 0} \frac{\|(A + B)\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\| + \|B\mathbf{x}\|}{\|\mathbf{x}\|}$$

$$\leqslant \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} + \sup_{\mathbf{x} \neq 0} \frac{\|B\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\| + \|B\|.$$

Thus, the triangle inequality for the norm of a matrix introduced by the formula (3) is fulfilled. The validity of the rest of the axioms for the norm (3) is obvious.

According to (3), for any $\mathbf{x} \neq 0$ we have

$$\|A\mathbf{x}\| \leqslant \|A\|\|\mathbf{x}\|, \tag{8}$$

and it is possible to prove that there exists such a vector $\mathbf{x} \neq 0$ for which the inequality (8) turns into an equality. Obviously, the relationship (8) turns into an equality for $\mathbf{x} = 0$ as well.

On the basis of (8), the following inequalities are fulfilled for any $\mathbf{x} \in \mathbf{R}^n$

$$\|AB\mathbf{x}\| = \|A(B\mathbf{x})\| \leqslant \|A\|\|B\mathbf{x}\| \leqslant \|A\|\|B\|\|\mathbf{x}\|$$

and, consequently,

$$\|A^2\mathbf{x}\| \leqslant \|A\|^2\|\mathbf{x}\|, \ldots, \|A^k\mathbf{x}\| \leqslant \|A\|^k\|\mathbf{x}\|,$$

i.e.

$$\|A^k\| \leqslant \|A\|^k. \tag{9}$$

Consider the system of linear algebraic equations

$$A\mathbf{x} = \mathbf{b}, \tag{10}$$

where $\det A \neq 0$ and $\mathbf{b} \neq 0$. The system (10) has the unique solution $\mathbf{x} \neq 0$. In practice, when solving the system (10) by any method, including the Gauss method, the relevant computations are performed by rounding

off,that is, incorrectly. The computational errors can often be interpreted as the errors of the right-hand side. Alongside with the system (10), let us consider the system

$$A(\mathbf{x} + \mathbf{r}) = \mathbf{b} + \eta, \tag{11}$$

where $\eta \neq 0$ is the error (perturbation) of the right-hand side, and $\mathbf{r}$ is the error of the solution. We have $A\mathbf{r} = \eta$ and $\mathbf{r} = A^{-1}\eta$,

$$\frac{\|\mathbf{r}\|/\|\mathbf{x}\|}{\|\eta\|/\|\mathbf{b}\|} = \frac{\|\mathbf{b}\|}{\|\mathbf{x}\|}\frac{\|\mathbf{r}\|}{\|\eta\|} = \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}\frac{\|A^{-1}\eta\|}{\|\eta\|}$$

$$\leqslant \frac{\|A\|\|\mathbf{x}\|}{\|\mathbf{x}\|}\frac{\|A^{-1}\|\|\eta\|}{\|\eta\|} = \|A\|\|A^{-1}\| = \nu(A),$$

that is, for the ratio of the relative error of the solution $\|\mathbf{r}\|/\|\mathbf{x}\|$ to the relative error of the right-hand side $\|\eta\|/\|\mathbf{b}\|$ the following inequality is fulfilled:

$$\frac{\|\mathbf{r}\|/\|\mathbf{x}\|}{\|\eta\|/\|\mathbf{b}\|} \leqslant \nu(A) = \|A\|\|A^{-1}\| \tag{12}$$

which turns into an equality for some $\mathbf{b} \neq \mathbf{0}$ and $\eta \neq \mathbf{0}$.

The quantity

$$\nu(A) = \|A\|\|A^{-1}\| \tag{13}$$

is called the *measure of the condition of matrix A. It is equal to the maximum possible enhancement ratio of the relative error of the right-hand member to the solution of the system* (10). If matrix A is symmetric and a second norm is chosen, that is, $\|A\| = \|A\|_2$, then, according to (7),

$$\nu(A) = \frac{\max\limits_{1 \leqslant i \leqslant n} |\lambda_i(A)|}{\min\limits_{1 \leqslant i \leqslant n} |\lambda_i(A)|}. \tag{14}$$

If $\nu(A)$ is large, then matrix A (the system (10)) is said to be *ill-conditioned,* and if $\nu(A)$ is small, then it is said to be *well-conditioned.*

An Example of an Ill-conditioned System of Equations. Consider the system of equations (10) in which $\mathbf{b} = (-1, -1, \ldots, -1, 1)$,

$$A = \begin{pmatrix} 1 & -1 & -1 & \ldots & -1 & -1 \\ 0 & 1 & -1 & \ldots & -1 & -1 \\ 0 & 0 & 1 & \ldots & -1 & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \tag{15}$$

and $\det A = 1 \neq 0$. In a more detailed notation of this system it is written as follows:

$$\begin{aligned} x_1 - x_2 - x_3 - \ldots - x_n &= -1, \\ x_2 - x_3 - \ldots - x_n &= -1, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ x_{n-1} - x_n &= -1, \\ x_n &= 1. \end{aligned} \tag{16}$$

The system (16) has the only solution $\mathbf{x} = (0, 0, \ldots, 0, 1)$. When the solution of the system (16) is found by the Gauss method, the forward procedure is actually absent. Suppose that in the backward procedure we committed a unique error which consisted in that $\tilde{x}_n = x_n + r_n = 1 + \varepsilon$ was obtained instead of $x_n = 1$, where $\varepsilon \neq 0$, is small as compared with unity. Then instead of the solution of the system (16), there will be found the solution $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{r}$ of the system (11), where matrix A and vector $\mathbf{b}$ have the above-indicated form, $\eta = (0, 0, \ldots, 0, \varepsilon)$. The error $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ satisfies the system of equations

$$\begin{aligned} r_1 - r_2 - r_3 - \ldots - r_n &= 0, \\ r_2 - r_3 - \ldots - r_n &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ r_{n-1} - r_n &= 0, \\ r_n &= \varepsilon. \end{aligned}$$

Hence we obtain

$$\begin{aligned} r_n &= \varepsilon, \\ r_{n-1} &= r_n = \varepsilon, \\ r_{n-2} &= r_n + r_{n-1} = \varepsilon + \varepsilon = 2\varepsilon, \\ r_{n-3} &= r_n + r_{n-1} + r_{n-2} = \varepsilon + \varepsilon + 2\varepsilon = 4\varepsilon = 2^2\varepsilon, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{n-k} &= r_n + r_{n-1} + \ldots + r_{n-(k-1)} = 2^{k-1}\varepsilon, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_1 &= r_{n-(n-1)} = 2^{(n-1)-1}\varepsilon = 2^{n-2}\varepsilon. \end{aligned}$$

We have $\|\mathbf{r}\|_1 = 2^{n-2}|\varepsilon|$, $\|\mathbf{x}\|_1 = 1$, $\|\eta\|_1 = |\varepsilon|$, $\|\mathbf{b}\|_1 = 1$ and, according to (12),

$$\nu(A) = \nu_1(A) = \|A\|_1\|A^{-1}\|_1 \geqslant \frac{\|\mathbf{r}\|_1/\|\mathbf{x}\|_1}{\|\eta\|_1/\|\mathbf{b}\|_1} = 2^{n-2}. \tag{17}$$

In accordance with (5), we find $\|A\|_1 = n$, where A is the matrix (15). Consequently, along with the measure of the condition of $\nu(A)$, the norm of the inverse matrix is large, that is, $\|A^{-1}\|_1$, although $\det A^{-1} = 1/\det A = 1$. For instance, if $n = 102$, then $\|A\|_1 = 102$ and, on the basis of (17), $\nu(A) \geqslant 2^{100} > 10^{30}$ and $\|A^{-1}\|_1 > 10^{27}$. In this case, $\|\mathbf{r}\|_1 = 2^{100}|\varepsilon| > 10^{30}|\varepsilon|$. In particular, if $|\varepsilon| = 10^{-15}$ (i.e. the only error committed in the backward procedure is sufficiently small), then the error of the solution obtained is nevertheless large: $\|\mathbf{r}\|_1 > 10^{15}$. This is accompanied by an ill condition of the system of equations established by the inequality (17).

An example of a well-conditioned system is given in the next section.

Sec. 21. The Method of Simple Iterations. Seidel's Method

The Method of Simple Iterations. Consider the system of linear equations of the form

$$\mathbf{x} = B\mathbf{x} + \mathbf{b}, \tag{1}$$

where B is a given square number matrix of order n, and $\mathbf{b} \in \mathbf{R}^n$ is a given vector (a constant term).

The *method of simple iterations* consists in the following. We choose an arbitrary vector $\mathbf{x}^0 \in \mathbf{R}^n$ (the initial approximation) and construct an iterative sequence of vectors using the formula

$$\mathbf{x}^k = B\mathbf{x}^{k-1} + \mathbf{b}, \quad k = 1, 2, \ldots . \tag{2}$$

Here, we give a theorem yielding the sufficient condition for the convergence of the method of simple iterations.

Theorem 1. *If* $\|B\| < 1$, *then the system of equations* (1) *has the unique solution* $\mathbf{x}^*$ *and the iterations (successive approximations)* (2) *converge to the solution with the rate of a geometric progression.*

Proof. Suppose that $\mathbf{x}^*$ is the solution of the system (1), that is, the following equality is fulfilled:

$$\mathbf{x}^* = B\mathbf{x}^* + \mathbf{b}. \tag{3}$$

Hence, in terms of the triangle inequality, for the norm of the vector and the inequality (20.8) we obtain

$$\|\mathbf{x}^*\| \leqslant \|B\mathbf{x}^*\| + \|\mathbf{b}\| \leqslant \|B\|\|\mathbf{x}^*\| + \|\mathbf{b}\|,$$

i.e. $(1 - \|B\|)\|\mathbf{x}^*\| \leqslant \|\mathbf{b}\|$ or, since $1 - \|B\| > 0$,

$$\|\mathbf{x}^*\| \leqslant \frac{\|\mathbf{b}\|}{1 - \|B\|}. \tag{4}$$

This inequality implies the uniqueness of the solution of the homogeneous system $\mathbf{x} = B\mathbf{x}$, i.e. for $\|\mathbf{b}\| = 0$, and consequently, the existence and uniqueness of the solution of the system (1) for any constant term $\mathbf{b}$.

Let $\mathbf{r}^k = \mathbf{x}^* - \mathbf{x}^k$, where $\mathbf{x}^*$ is the solution of the system (1). Subtracting the equality (2) from the equality (3), we find

$$\mathbf{x}^* - \mathbf{x}^k = B(\mathbf{x}^* - \mathbf{x}^{k-1}), \tag{5}$$

that is, $\mathbf{r}^k = B\mathbf{r}^{k-1}$ and, consequently, $\mathbf{r}^k = B^k\mathbf{r}^0$, where B^k is the kth power of matrix B, $\mathbf{r}^k$ is the error of the kth iteration, and $\mathbf{r}^0$ is the initial error. Hence, on the basis of (20.8) and (20.9), we obtain

$$\|\mathbf{r}^k\| \leqslant \|B^k\|\|\mathbf{r}^0\| \leqslant \|B\|^k\|\mathbf{r}^0\|,$$

that is, the norm of the error $\mathbf{r}^k = \mathbf{x}^* - \mathbf{x}^k$ tends to zero as $k \to \infty$ not slower than the geometric progression with the denominator $q = \|B\| < 1$.

Theorem 1 has been proved completely.

Remark 1. Let r_i^k be the ith component of the error $\mathbf{r}^k$ of the kth iteration (kth approximation) of $\mathbf{x}^k$. Since $|r_i^k| \leqslant \|r^k\|$, $i = 1, 2, \ldots, n$, all the components r_i^k tend to zero as $k \to \infty$ not slower than $\|B\|^k\|\mathbf{r}^0\|$.

Remark 2. By virtue of the inequality (20.2), from the convergence of the iterations with respect to one of the two introduced norms of the vectors there follows the convergence with respect to the other norm, that is, if $\|\mathbf{r}^k\|_1 = \|\mathbf{x}^* - \mathbf{x}^k\|_1 \to 0$ as $k \to \infty$, then $\|\mathbf{r}^k\|_2 = \|\mathbf{x}^* - \mathbf{x}^k\|_2 \to 0$ as $k \to \infty$, and vice versa.

Estimating the Error of Iterations. Let $\|B\| < 1$. Then by Theorem 1, the system (1) has the unique solution $\mathbf{x}^*$ for which the equality (3) is fulfilled. From the equalities (2) and (3) there follows the equality

$$\mathbf{x}^* - \mathbf{x}^{k-1} = \mathbf{x}^k - \mathbf{x}^{k-1} + B(\mathbf{x}^* - \mathbf{x}^{k-1}).$$

Hence we get

$$\|\mathbf{x}^* - \mathbf{x}^{k-1}\| \leqslant \|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \|B(\mathbf{x}^* - \mathbf{x}^{k-1})\|$$
$$\leqslant \|\mathbf{x}^k - \mathbf{x}^{k-1}\| + \|B\|\|\mathbf{x}^* - \mathbf{x}^{k-1}\|$$

or $(1 - \|B\|)\|\mathbf{x}^* - \mathbf{x}^{k-1}\| \leqslant \|\mathbf{x}^k - \mathbf{x}^{k-1}\|$, and since $1 - \|B\| > 0$, we get

$$\|\mathbf{x}^* - \mathbf{x}^{k-1}\| \leqslant \frac{1}{1 - \|B\|}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|. \tag{6}$$

Besides, by virtue of (5), we have

$$\|\mathbf{x}^* - \mathbf{x}^k\| \leqslant \|B\|\|\mathbf{x}^* - \mathbf{x}^{k-1}\|. \tag{7}$$

From (6) and (7) we finally obtain

$$\|\mathbf{x}^* - \mathbf{x}^k\| \leqslant \frac{\|B\|}{1 - \|B\|} \|\mathbf{x}^k - \mathbf{x}^{k-1}\|. \tag{8}$$

The inequality (8) makes it possible to estimate the norm of the error of the kth approximation in terms of the norm of the difference of two successive approximations and the norm of matrix B. This estimate is widely used in practice.

To compute the norm $\|B\|$ or to obtain its estimate, we use the formulas (20.5)-(20.7).

Examples:

(1) $$B = \begin{pmatrix} -\frac{3}{5} & \frac{3}{5} \\ \frac{2}{5} & \frac{1}{5} \end{pmatrix},$$

$$\|B\|_1 = \max_{1 \leqslant i \leqslant 2} \sum_{j=1}^{2} |b_{ij}| = \frac{6}{5} > 1,$$

$$\|B\|_2 \leqslant \left(\sum_{i,j=1}^{2} b_{ij}^2 \right)^{1/2} = \frac{\sqrt{23}}{5} < 1;$$

(2) $$B = \begin{pmatrix} \frac{1}{10} & \frac{4}{5} \\ \frac{3}{5} & -\frac{1}{5} \end{pmatrix},$$

$$\|B\|_1 = \frac{9}{10} < 1, \quad \|B\|_2 \leqslant \left(\sum_{i,j=1}^{2} b_{ij}^2 \right) = \sqrt{\frac{21}{20}},$$

that is, this estimate is of no interest for $\|B\|_2$, since the right-hand member is greater than unity.

By virtue of Remark 2, in order for the iterations (2) to be convergent, it is sufficient that at least one of the norms of matrix B be less than unity.

Let us set forth (without proof) a theorem which provides the necessary and sufficient condition for the convergence of the method of simple iterations.

Theorem 2. *Let the system* (1) *have a unique solution. The successive approximations* (2) *converge to the solution of the system* (1) *for any initial approximation* $\mathbf{x}^0$ *if and only if all the eigenvalues of matrix B are less than unity (by the absolute value).*

Theorem 2 gives more general conditions for the convergence of the method of simple iterations as compared with Theorem 1, since it may happen that the conditions of Theorem 2 are fulfilled, whereas those of Theorem 1 are not fulfilled. But, in the general case, it is not a simple thing to make use of Theorem 2, since a knowledge of the boundaries of the eigenvalues of matrix B is required. In a particular case, when the matrix is symmetric,given in Sec. 23 is the method of finding the maximum and minimum eigenvalues of a matrix which makes it possible to verify the conditions of Theorem 2 and, in particular, to compute the second norm of the matrix by the formula (20.7).

An Example of a Well-conditioned System of Equations. The system of equations (1) is equivalent to the system of equations $\mathbf{x} - B\mathbf{x} = \mathbf{b}$ or, which is the same, $A\mathbf{x} = \mathbf{b}$, where $A = E - B$, E being a unit matrix. Under the condition that $\|B\| < 1$, according to Theorem 1, $\det A \neq 0$, and, consequently, $\mathbf{x}^* = A^{-1}\mathbf{b}$ is the solution of the system (1). Hence

$$\|\mathbf{x}^*\| \leqslant \|A^{-1}\|\|\mathbf{b}\|, \tag{9}$$

and there exists a constant term $\mathbf{b} \neq \mathbf{0}$ for which the inequality (9) turns into an equality. This and the inequality (4) imply that if $\|B\| < 1$, then

$$\|A^{-1}\| \leqslant \frac{1}{1 - \|B\|}.$$

Taking into consideration (20.4), we get

$$\|A\| = \|E - B\| \leqslant \|E\| + \|B\| = 1 + \|B\|.$$

Thus, in accordance with (20.13), the measure of the condition of the matrix $A = E - B$ corresponding to the system (1) satisfies the inequality if $\|B\| < 1$

$$\nu(A) = \nu(E - B) = \|A\|\|A^{-1}\| \leqslant \frac{1 + \|B\|}{1 - \|B\|}. \tag{10}$$

For instance, let the following numbers be elements of matrix B of order n:

$$b_{ij} = \frac{0.8}{n}(-1)^{i+j},\ 0 \leqslant i, j \leqslant n.$$

Then

$$\|B\|_1 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |b_{ij}| = \sum_{j=1}^{n} \frac{0.8}{n} = 0.8,$$

$$\|B\|_2 \leqslant \left(\sum_{i,j=1}^{n} b_{ij}^2 \right)^{1/2} = \left(\sum_{i,j=1}^{n} \left(\frac{0.8}{n} (-1)^{i+j} \right)^2 \right)^{1/2}$$

$$= \left(n^2 \left(\frac{0.8}{n} \right)^2 \right)^{1/2} = 0.8.$$

Thus, according to (10),

$$\nu(A) = \nu(E - B) \leqslant \frac{1 + 0.8}{1 - 0.8} = 9,$$

that is, the measure of the condition of matrix A is not large for any n. The corresponding system (1) is well-conditioned, and consequently, its solution will not be highly sensitive to the perturbations of the constant term $\mathbf{b}$, unlike, for instance, the system (20.10) with the matrix (20.15).

Seidel's Method. Let there be given the system of linear equations

$$A\mathbf{x} = \mathbf{b}, \tag{11}$$

where all the diagonal elements of the matrix $A = (a_{ij})_{i,j=1}^{n}$ are different from zero, that is, $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$. Dividing the ith equation of the system (11), $i = 1, 2, \ldots, n$, by a_{ii} and then transposing all the unknowns, except for x_i, to the right, we shall arrive at an equivalent system of the form

$$\mathbf{x} = C\mathbf{x} + \mathbf{d}, \tag{12}$$

where $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, $d_i = b_i/a_{ii}$ and $C = (c_{ij})_{i,j=1}^{n}$,

$$c_{ij} = \begin{cases} -a_{ij}/a_{ii}, & j \neq i, \\ 0, & j = i. \end{cases}$$

Seidel's method consists in that iterations are carried out by the formula

$$x_i^k = \sum_{j=1}^{i-1} c_{ij} x_j^k + \sum_{j=i+1}^{n} c_{ij} x_j^{k-1} + d_i, \tag{13}$$

where x_i^0 are arbitrary, $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots$.

The iterations (13) realized by Seidel's method differ from the simple iterations (2) in that we at once use the already found components of the kth approximation with lesser numbers when finding the ith component of the kth approximation.

The conditions for the convergence of the method of simple iterations and Seidel's method do not coincide, but intersect. In some cases, Seidel's method yields a more rapid convergence. Let us formulate the theorem on two different sufficient conditions for the convergence of Seidel's method.

Theorem 3. *For the unique solution of the system* (11) *to exist and Seidel's method to converge, it is sufficient that at least one of the following two conditions be fulfilled:*

(a) $\sum_{j \neq i} |a_{ij}| < |a_{ii}|, \ i = 1, 2, \ldots, n;$

(b) *matrix A is symmetric and positive definite (all of its eigenvalues are positive).*

Remark 3. The condition (b) can be verified by the method of finding the minimum eigenvalue of a symmetric matrix given in Sec. 23.

Sec. 22. The Sweep Method

Consider the system of linear algebraic equations of the following special form:

$$A_j z_{j-1} - C_j z_j + B_j z_{j+1} = F_j, \quad j = 1, 2, \ldots, N-1, \tag{1}$$

$$z_0 = \varkappa_0 z_1 + \nu_0, \quad z_N = \varkappa_N z_{N-1} + \nu_N, \tag{2}$$

where $z_0, z_1, \ldots, z_N$ are unknowns, $A_j, B_j, C_j, F_j, \varkappa_i, \nu_i$ are given numbers, and

$$|C_j| \geqslant |A_j| + |B_j| \geqslant |A_j| > 0, \quad |\varkappa_0| < 1, \quad |\varkappa_N| \leqslant 1. \tag{3}$$

The system (1), (2) can also be written in vector form: $A\mathbf{z} = \mathbf{b}$, where $\mathbf{z} = (z_0, z_1, \ldots, z_N)$ and $\mathbf{b} = (\nu_0, F_1, F_2, \ldots, F_{N-1}, \nu_N)$,

$$A = \begin{pmatrix} 1 & -\varkappa_0 & & & & & \\ A_1 & -C_1 & B_1 & & & 0 & \\ & A_2 & -C_2 & B_2 & & & \\ & & & & & & \\ & & & & & & \\ & 0 & & A_{N-1} & -C_{N-1} & B_{N-1} & \\ & & & & -\varkappa_N & 1 & \end{pmatrix}.$$

Thus, matrix A of the system (1), (2) is of the *tridiagonal* type, that is, all of its elements which are not in the principal diagonal and two neighbouring diagonals are zero. This is explained by the fact that each equation of the system (1), (2) contains only those unknowns whose numbers differ from the number of the equation by not more than unity. With the conditions (3) fulfilled, the matrix under consideration is said to be a matrix with the *dominating (prevailing) principal diagonal.*

The set of the equations (1) is usually called a *difference equation of second order,* or a *three-point difference equation,* and the equations (2) are termed the *boundary conditions* for the difference equation (1).

The problem (1), (2) as a whole is called the *boundary-value problem for a three-point difference equation,* or simply the *difference boundary-value problem.*

Difference boundary-value problems (systems of linear algebraic equations) of the form (1), (2) are widely encountered. For instance, the system of equations obtained in Sec. 11 when constructing the spline has the form (1), (2) and satisfies the conditions (3). Difference boundary-value problems occur most frequently in the approximate solution of differential equations (see Secs. 29, 31, and 33).

The conditions (3) guarantee the existence of the unique solution of the difference boundary-value problem (1), (2) (we shall make sure of this later on) and allow us to find this solution by using a special economical method, called the *sweep method.*

The Sweep Method. Substituting $z_0 = \varkappa_0 z_1 + \nu_0$ from (2) into the first equation of the system (1), we get

$$A_1(\varkappa_0 z_1 + \nu_0) - C_1 z_1 + B_1 z_2 = F_1$$

or

$$z_1 = \varkappa_1 z_2 + \nu_1, \tag{4}$$

where

$$\varkappa_1 = \frac{B_1}{C_1 - A_1\varkappa_0}, \quad \nu_1 = \frac{A_1\nu_0 - F_1}{C_1 - A_1\varkappa_0}.$$

Substituting the expression (4) for z_1 into the next equation of the system (1), we obtain the equation relating z_2 and z_3, and so forth. Suppose that we have already found the relationship

$$z_{k-1} = \varkappa_{k-1} z_k + \nu_{k-1}, \quad k < N - 1. \tag{5}$$

Substituting the expression (5) for z_{k-1} into the kth equation of system (1), we obtain

$$A_k(\varkappa_{k-1}z_k + \nu_{k-1}) - C_k z_k + B_k z_{k+1} = F_k.$$

Solving this equation with respect to z_k, we have

$$z_k = \varkappa_k z_{k+1} + \nu_k, \tag{6}$$

where

$$\varkappa_k = \frac{B_k}{C_k - A_k\varkappa_{k-1}}, \quad \nu_k = \frac{A_k\nu_{k-1} - F_k}{C_k - A_k\varkappa_{k-1}}. \tag{7}$$

Thus, the coefficients of the equations (6) relating the neighbouring values z_k and z_{k+1}, $k = 1, 2, \ldots, N-1$, can be determined by the recurrence relationships (7), since $\varkappa_0$ and ν_0 are specified in (2).

Substituting the expression for z_{N-1} which follows from the formula (6) for $k = N-1$ into the second boundary condition (2), we obtain

$$z_N = \varkappa_N(\varkappa_{N-1}z_N + \nu_{N-1}) + \nu_N, \tag{8}$$

where $\varkappa_N$, ν_N are the coefficients given in (2), and $\varkappa_{N-1}$, ν_{N-1} are computed by the formulas (7). From the equation (8) we find the unknown z_N:

$$z_N = \frac{\nu_N + \varkappa_N\nu_{N-1}}{1 - \varkappa_N\varkappa_{N-1}}. \tag{9}$$

Then using the formula (6), we find the remaining unknowns z_{N-1}, $z_{N-2}, \ldots, z_0$ in the reverse order. For $k = 0$ the formula (6) coincides with the first boundary condition (2).

The process of computing the coefficients $\varkappa_k$ and ν_k, $k = 1, 2, \ldots, N-1$, by the formulas (7) is called the *forward procedure of sweeping*, and finding the unknowns z_m, $m = N, N-1, \ldots, 0$ by the formulas (9) and (6) is called the *backward procedure of sweeping*.

By virtue of the conditions (3), the computations in terms of the formulas (7) and (9) are correct, that is, the denominators do not vanish in them. Let us verify this.

Suppose that for some k, $0 < k < N-1$, the inequality $|\varkappa_{k-1}| < 1$ is fulfilled. For instance, under one of the conditions (3) this inequality is fulfilled for $k = 1$. Since $|A_k| > 0$, making use of the

conditions (3), we obtain

$$|C_k - A_k\varkappa_{k-1}| \geqslant |C_k| - |A_k||\varkappa_{k-1}| > |C_k| - |A_k| \geqslant 0,$$

$$|\varkappa_k| = \frac{|B_k|}{|C_k - A_k\varkappa_{k-1}|} \leqslant \frac{|C_k| - |A_k|}{|C_k| - |A_k||\varkappa_{k-1}|} < 1.$$

Hence, by induction, it follows that

$$|\varkappa_k| < 1, \quad k = 0, 1, \ldots, N-1, \tag{10}$$

and, consequently, the denominators in the expressions (7) are not zero.

Since $|\varkappa_N| \leqslant 1$ (by the right condition (3)) and $|\varkappa_{N-1}| < 1$ (by proof), we have $1 - \varkappa_N\varkappa_{N-1} > 0$, that is, the denominator in the formula (9) is not equal to zero.

Lemma 1. *For the conditions* (3), *the difference boundary-value problem* (1), (2) *has a unique solution.*

Proof. Having obtained the explicit formulas (7), (9), and (6) for finding the solution of the problem (1), (2) under the conditions (3), we have thus established its solvability. The uniqueness of the solution follows from the solvability for any constant terms ν_1, ν_N, and F_j, $j = 1, 2, \ldots, N-1$. Indeed, if the determinant of a system of linear algebraic equations is equal to zero, then, according to the Kronecker-Capelli theory, such a system has a solution not for any constant terms. Consequently, the determinant is not zero in the system (1), (2), therefore it has a unique solution.

Lemma 1 has been proved.

Let us count the number of the arithmetic operations to be performed when solving the difference boundary-value problem (1), (2) in terms of the sweep method. By the formulas (7) realized with the aid of six arithmetic operations, computations are performed $N-1$ times, $k = 1, 2, \ldots, N-1$, by the formula (9) five arithmetic operations are performed, and, finally, by the formula (6) requiring only two operations, computations are realized N times, $k = N-1, N-2, \ldots, 0$. Thus, in the sweep method the total of arithmetic operations amounts to

$$Q = 6(N-1) + 5 + 2N = 8(N+1) - 9, \tag{11}$$

that is, the number of operations grows linearly with respect to the number of the unknowns equal to $N+1$. And when solving an arbitrary system of linear algebraic equations by the Gauss method, the number of operations expressed by the formula (19.18) is approximately proportional to the cube of the number of the unknowns.

Let us prove two lemmas yielding the estimates of the solution of the problem (1), (2) in two particular cases. These estimates are used in Sec. 29.

Lemma 2. *If* $\varkappa_0 = \varkappa_N = \nu_0 = \nu_N = 0$ *and* $A_j = B_j = 1$, $C_j \geqslant 2$, $j = 1, 2, \ldots, N-1$, *then the solution* $\mathbf{z} = (z_0, z_1, \ldots, z_N)$ *of the difference boundary-value problem* (1), (2) *satisfies the inequality*

$$\max_{0 \leqslant j \leqslant N} |z_j| \leqslant N^2 \max_{1 \leqslant j \leqslant N-1} |F_j|. \tag{12}$$

Proof. The conditions (3) are fulfilled. Consequently, by Lemma 1, the problem (1), (2) has the unique solution $\mathbf{z} = (z_0, z_1, \ldots, z_N)$, where, by virtue of (2), $z_0 = z_N = 0$. And the problem (1), (2) can be solved by the sweep method. According to (10), $|\varkappa_k| < 1$, $k = 0, 1, \ldots, N-1$. Consequently, for $k = 1, 2, \ldots, N-1$

$$0 < \frac{A_k}{C_k - A_k \varkappa_{k-1}} = \frac{1}{C_k - A_k \varkappa_{k-1}} \leqslant \frac{1}{2 - \varkappa_{k-1}} < 1.$$

Therefore, in accordance with (7),

$$|\nu_k| \leqslant |\nu_{k-1}| + |F_k|, \quad k = 1, 2, \ldots, N-1.$$

Since $\nu_0 = 0$,

$$|\nu_k| \leqslant \sum_{i=1}^{k} |F_i|, \quad k = 1, 2, \ldots, N-1,$$

and hence

$$\max_{1 \leqslant j \leqslant N-1} |\nu_j| \leqslant \sum_{i=1}^{N-1} |F_i| \leqslant N \max_{1 \leqslant j \leqslant N-1} |F_j|.$$

Combining this with (6) and taking into consideration that $|\varkappa_k| < 1$, $k = N-1, N-2, \ldots, 1$, and $z_0 = z_N = 0$, we obtain the inequality (12) in an analogous manner.

Lemma 3. *If* $\varkappa_0 = \varkappa_N = 0$ *and* $F_j = 0$, $A_j = B_j = 1$, $C_j \geqslant 2$, $j = 1, 2, \ldots, N-1$, *then the solution of the difference boundary-value problem* (1), (2) *satisfies the relationship*

$$\max_{0 \leqslant j \leqslant N} |z_j| = \max\{|\nu_0|, |\nu_N|\}. \tag{13}$$

Proof. By virtue of the conditions (3) in accordance with Lemma 1, the solution $\mathbf{z} = (z_0, z_1, \ldots, z_N)$ of the problem (1), (2) exists and is unique. Obviously, the following equality is valid for some k:

$$|z_k| = \max_{0 \leqslant j \leqslant N} |z_j|. \tag{14}$$

If $k = 0$ or N, then from this circumstance and from (2), where $\varkappa_0 = \varkappa_N = 0$, the relationship (13) follows directly.

Suppose that $0 < k < N$. Then, according to (1) under the conditions of Lemma 3, we have

$$z_{k-1} - C_k z_k + z_{k+1} = 0, \tag{15}$$

where $C_k \geqslant 2$. By virtue of (14), the equality (15) is possible for $C_k = 2$ only if $z_{k+1} = z_k = z_{k-1}$, and for $C_k > 2$ from (14) and (15) it necessarily follows that $z_{k+1} = z_k = z_{k-1} = 0$. Thus, along with (14), we have

$$|z_{k+1}| = \max_{0 \leqslant j \leqslant N} |z_j|.$$

Similarly, we make sure that $z_{k+2} = z_{k+1} = z_k$, and so forth. Over a finite number of steps we shall come to a conclusion that $z_N = z_k$. Combining this with (14) and (2), where $\varkappa_0 = \varkappa_N = 0$, yields the equality (13).

Remark 1. The property of the solution of the problem (1), (2) (under the conditions of Lemma 3) expressed by the equality (13) is called the *maximum-modulus principle.* According to this principle, the solution attains a maximum-modulus value at least at one of the extreme points (nodes), since $z_0 = \nu_0$ and $z_N = \nu_N$.

Sec. 23. Partial Eigenvalue Problems

The problem of finding all eigenvalues and eigenvectors of a number matrix is referred to as a *complete eigenvalue problem.* In the general case, this problem is rather complicated. We shall dwell on some important particular problems only.

Finding the Maximum-modulus Eigenvalue. Suppose that a square matrix A of order n has a complete system of normed eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, that is, in particular,

$$A\mathbf{e}_i = \lambda_i \mathbf{e}_i, \tag{1}$$

$$\|\mathbf{e}_i\| = \|\mathbf{e}_i\|_2 = (\mathbf{e}_i, \mathbf{e}_i)^{1/2} = 1, \tag{2}$$

where $(\mathbf{x}, \mathbf{y})$ is a scalar product in n-dimensional Euclidean vector space, and λ_i is the eigenvalue of matrix A corresponding to the eigenvector $\mathbf{e}_i$, $i = 1, 2, \ldots, n$. For instance, a complete system of eigenvectors exists if matrix A is symmetric.

Suppose that

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|. \tag{3}$$

Let us take an arbitrary vector $\mathbf{x}^0 \neq 0$. We have

$$\mathbf{x}^0 = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_n\mathbf{e}_n,$$

where $c_1, c_2, \ldots, c_n$ are the coordinates of the vector $\mathbf{x}^0$ in the basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

Suppose that

$$c_1 \neq 0. \tag{4}$$

In turn, we find the vectors

$$\mathbf{x}^k = A\mathbf{x}^{k-1}, \quad k = 1, 2, \ldots . \tag{5}$$

Then, according to (1),

$$\begin{aligned}\mathbf{x}^1 = A\mathbf{x}^0 &= A(c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_n\mathbf{e}_n) \\ &= c_1A\mathbf{e}_1 + c_2A\mathbf{e}_2 + \ldots + c_nA\mathbf{e}_n \\ &= c_1\lambda_1\mathbf{e}_1 + c_2\lambda_2\mathbf{e}_2 + \ldots + c_n\lambda_n\mathbf{e}_n.\end{aligned}$$

And, in general,

$$\mathbf{x}^k = c_1\lambda_1^k\mathbf{e}_1 + c_2\lambda_2^k\mathbf{e}_2 + \ldots + c_n\lambda_n^k\mathbf{e}_n = \lambda_1^k(c_1\mathbf{e}_1 + \eta^k), \tag{6}$$

where

$$\eta^k = c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{e}_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathbf{e}_n,$$

and, by virtue of (3),

$$\|\eta^k\| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \to 0 \text{ as } k \to \infty. \tag{7}$$

Here and below in the present section, the superscript of a vector, for instance, $\mathbf{x}^k$, indicates the number of approximation, in scalars, for instance, λ_1^k, the superscript means the power, and the second norm is taken as the norm of vectors, that is, $\|\mathbf{x}\| = \|\mathbf{x}\|_2 = (\mathbf{x}, \mathbf{x})^{1/2}$.

Taking into account (6), we obtain the following expression for the scalar product $(\mathbf{x}^k, \mathbf{x}^{k-1})$:

$$\begin{aligned}(\mathbf{x}^k, \mathbf{x}^{k-1}) &= \lambda_1^{2k-1}(c_1\mathbf{e}_1 + \eta^k, c_1\mathbf{e}_1 + \eta^{k-1}) \\ &= \lambda_1^{2k-1}[c_1^2(\mathbf{e}_1, \mathbf{e}_1) + c_1(\mathbf{e}_1, \eta^{k-1}) \\ &\qquad + c_1(\eta^k, \mathbf{e}_1) + (\eta^k, \eta^{k-1})].\end{aligned} \tag{8}$$

Since, according to (2), $(\mathbf{e}_1, \mathbf{e}_1) = \|\mathbf{e}_1\|^2 = 1$ and by Buniakowski's inequality*

$$\begin{aligned} |(\mathbf{e}_1, \eta^{k-1})| &\leqslant \|\mathbf{e}_1\| \|\eta^{k-1}\| = \|\eta^{k-1}\|, \\ |(\eta^k, \mathbf{e}_1)| &\leqslant \|\eta^k\| \|\mathbf{e}_1\| = \|\eta^k\|, \\ |(\eta^k, \eta^{k-1})| &\leqslant \|\eta^k\| \|\eta^{k-1}\|, \end{aligned}$$

from (8), bearing in mind (7), we find

$$(\mathbf{x}^k, \mathbf{x}^{k-1}) = \lambda_1^{2k-1} \left(c_1^2 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{k-1} \right) \right).$$

Analogously, we obtain

$$(\mathbf{x}^m, \mathbf{x}^m) = \lambda_1^{2m} \left(c_1^2 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^m \right) \right).$$

Consequently,

$$\lambda_{1k} = \frac{(\mathbf{x}^k, \mathbf{x}^{k-1})}{(\mathbf{x}^{k-1}, \mathbf{x}^{k-1})} = \lambda_1 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{k-1} \right), \tag{9}$$

$$\|\mathbf{x}^k\| = (\mathbf{x}^k, \mathbf{x}^k)^{1/2} = |\lambda_1|^k \left(|c_1| + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right), \tag{10}$$

$$\mathbf{e}_1^k = \frac{\mathbf{x}^k}{\|\mathbf{x}^k\|} = (\operatorname{sgn} \lambda_1)^k \mathbf{e}_1' + \mathbf{r}^k, \tag{11}$$

where $\mathbf{e}_1' = (\operatorname{sgn} c_1)\mathbf{e}_1$ and $\|\mathbf{r}^k\| = O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right)$.

Thus, with the condition (3) the iteration process (5) makes it possible to find with any accuracy the maximum-modulus eigenvalue λ_1 (since $\lambda_{1k} \to \lambda_1$ as $k \to \infty$) and its corresponding eigenvector $\mathbf{e}_1'$ (since $\|\mathbf{r}^k\| \to 0$ as $k \to \infty$).

In practice the scalar products $(\mathbf{x}^k, \mathbf{x}^{k-1})$, $(\mathbf{x}^{k-1}, \mathbf{x}^{k-1})$, $(\mathbf{x}^k, \mathbf{x}^k)$ entering into the formulas (9) and (10) are computed for the vectors $\mathbf{x}^k$ and $\mathbf{x}^{k-1}$ represented in the original orthonormal basis in which operator A is specified, that is, by the usual formula

$$(\mathbf{x}, \mathbf{y}) = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n.$$

---

* Buniakowski's inequality $|(\mathbf{x}, \mathbf{y})| \leqslant \|\mathbf{x}\| \|\mathbf{y}\|$ is provided in the course of linear algebra.

Remark 1. If $|\lambda_1| > 1$, then $\|\mathbf{x}^k\| \to \infty$, and if $|\lambda_1| < 1$, then $\|\mathbf{x}^k\| \to 0$ as $k \to \infty$. Both of these phenomena are undesirable when carrying out calculations on a computer. In the first case, overfilling may occur with the computer, which results in computation coming to a stop. In the second case, $\|\mathbf{x}^k\|$ may become a computer zero (too small quantity) and the desired information is lost. Therefore it is advisable to carry out iterations by the formulas

$$\mathbf{e}_1^0 = \frac{\mathbf{x}^0}{\|\mathbf{x}^0\|},$$

$$\mathbf{x}^k = A\mathbf{e}_1^{k-1}, \quad \lambda_{1k} = (\mathbf{x}^k, \mathbf{e}_1^{k-1}), \quad \mathbf{e}_1^k = \frac{\mathbf{x}^k}{\|\mathbf{x}^k\|} \tag{9*}$$

which are free of the indicated disadvantage and yield the same result (the same $\mathbf{e}_1^k$ and λ_{1k}) as the formulas (5), (9), and (11).

Remark 2. If when choosing $\mathbf{x}^0$, the condition (4) is not fulfilled, which is unlikely, then owing to round-off errors a nonzero component corresponding to $\mathbf{e}_1$ will appear,as a rule, after several iterations. Then, although with some delay, the iteration process will yield the first eigenvalue.

Remark 3. That λ_1 is not a multiple eigenvalue and that there is no eigenvalue equal to $-\lambda_1$ is confirmed by the convergence of the iteration process when choosing different $\mathbf{x}^0$ for one and the same eigenvector (within the opposite vector).

Let us now consider the case when matrix A is symmetric. Then, as is known, all of its eigenvalues are automatically real, and there exists an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ made up of the eigenvectors of matrix A.

Let E be a unit matrix, let

$$P_2(t) = a_0 + a_1 t + a_2 t^2$$

be some second-degree polynomial in t with real coefficients, and let B be a matrix expressed in terms of matrix A in the following way:

$$B = a_0 E + a_1 A + a_2 A^2.$$

It is easily seen that the eigenvalues of matrices A and B are connected by the relationship

$$\lambda_i(B) = P_2(\lambda_i(A)), \tag{12}$$

and the eigenvectors of matrix A corresponding to $\lambda_i(A)$ are eigenvectors of matrix B corresponding to the eigenvalue $\lambda_i(B)$.

Indeed, let $\mathbf{e}_i$ be the eigenvector of matrix A corresponding to the eigenvalue $\lambda_i(A)$, that is, $A\mathbf{e}_i = \lambda_i(A)\mathbf{e}_i$. Then

$$A^2\mathbf{e}_i = A(A\mathbf{e}_i) = A\lambda_i(A)\mathbf{e}_i = \lambda_i(A)A\mathbf{e}_i = \lambda_i^2(A)\mathbf{e}_i$$

and, consequently,

$$\begin{aligned} B\mathbf{e}_i &= (a_0E + a_1A + a_2A^2)\mathbf{e}_i \\ &= a_0E\mathbf{e}_i + a_1A\mathbf{e}_i + a_2A^2\mathbf{e}_i \\ &= a_0\mathbf{e}_i + a_1\lambda_i(A)\mathbf{e}_i + a_2\lambda_i^2(A)\mathbf{e}_i \\ &= (a_0 + a_1\lambda_i(A) + a_2\lambda_i^2(A))\mathbf{e}_i = P_2(\lambda_i(A))\mathbf{e}_i. \end{aligned}$$

Let us dwell on three more problems.

Finding λ_2 and $\mathbf{e}_2$. Let

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|,$$

λ_1 and $\mathbf{e}_1$ being known, and $\|\mathbf{e}_1\| = 1$.

Let us assign an iteration process:

$\mathbf{x}^0$ is an arbitrary vector $((\mathbf{x}^0, \mathbf{e}_2) \neq 0)$,

$\mathbf{y}^0 = \mathbf{x}^0 - (\mathbf{x}^0, \mathbf{e}_1)\mathbf{e}_1$, $\mathbf{e}_2^0 = \mathbf{y}^0/\|\mathbf{y}^0\|$,

$\mathbf{x}^k = A\mathbf{e}_2^{k-1}$,

$\lambda_{2k} = (\mathbf{x}^k, \mathbf{e}_2^{k-1})$,

$\mathbf{y}^k = \mathbf{x}^k - (\mathbf{x}^k, \mathbf{e}_1)\mathbf{e}_1$ (orthogonalization to $\mathbf{e}_1$),

$\mathbf{e}_2^k = \mathbf{y}^k/\|\mathbf{y}^k\|$ (normalization),

$k = 1, 2, \ldots$.

The same as above, taking an additional account of the orthonormalization of the basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, it is possible to prove that

$$\lambda_{2k} = \lambda_2 + O\left(\left|\frac{\lambda_3}{\lambda_2}\right|^{2k-1}\right),$$

$$\mathbf{e}_2^k = (\operatorname{sgn}\lambda_2)^k\mathbf{e}_2' + \boldsymbol{\rho}^k,$$

where $\mathbf{e}_2' = \mathbf{e}_2$ or $-\mathbf{e}_2$ and $\|\boldsymbol{\rho}^k\| = O\left(\left|\frac{\lambda_3}{\lambda_2}\right|^{2k}\right)$.

Finding $\max\limits_{1 \leqslant i \leqslant n} \lambda_i(A)$ and $\min\limits_{1 \leqslant i \leqslant n} \lambda_i(A)$. Let us assume that the maximum-modulus eigenvalue $\bar{\lambda}(A)$ of matrix A is known. We find the

maximum-modulus eigenvalue $\bar{\lambda}(B)$ of the matrix

$$B = A - \bar{\lambda}(A)E \tag{13}$$

using the method set forth at the beginning of this section.

If $\bar{\lambda}(A) > 0$, then, obviously,

$$\max_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A). \tag{14}$$

Besides, according to (12) and (13), $\lambda_i(B) = \lambda_i(A) - \bar{\lambda}(A) \leqslant 0 \; \forall \, i$. Therefore

$$\bar{\lambda}(B) = \min_{1 \leqslant i \leqslant n} (\lambda_i(A) - \bar{\lambda}(A)) = \min_{1 \leqslant i \leqslant n} \lambda_i(A) - \bar{\lambda}(A),$$

that is,

$$\min_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A) + \bar{\lambda}(B). \tag{15}$$

If $\bar{\lambda}(A) < 0$, then

$$\min_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A)$$

and, according to (12) and (13),

$$\max_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A) + \bar{\lambda}(B). \tag{16}$$

Remark 4. It is usually unknown beforehand whether a symmetric matrix A has maximum-modulus eigenvalues with opposite signs or a multiple maximum-modulus eigenvalue. These cases can be explained on the basis of the computation in the following way. If for different initial vectors $\mathbf{x}^0$ the eigenvalues λ_{1k} computed by the formulas (9*) converge not to one and the same value, then this testifies to the presence of maximum-modulus eigenvalues with opposite signs. Then we have to displace the spectrum taking the matrix $A' = A + cE$, where $c \neq 0$ is a number. If the eigenvalues λ_{1k} converge to one and the same number for different initial vectors $\mathbf{x}^0$, but the sequence of vectors $\mathbf{e}_1^k$ obtained by the formulas (9*) leads to noncollinear vectors $\mathbf{e}_1'$, then this confirms that the maximum-modulus eigenvalue is multiple. Here, λ_{1k} converges to the indicated eigenvalue, and the vectors $\mathbf{e}_1'$ thus obtained are eigenvectors.

Finding the Distance ρ_0 from the Given Point $\lambda = \lambda_0$ to the Nearest Eigenvalue of Matrix A. This problem arises when studying phenomena of the resonance type. Consider the case of the greatest interest:

$$\min_{1 \leqslant i \leqslant n} \lambda_i(A) < \lambda_0 < \max_{1 \leqslant i \leqslant n} \lambda_i(A).$$

Let us set

$$\rho_0 = \min_{1 \leqslant i \leqslant n} |\lambda_i(A) - \lambda_0|,$$

$$l = \max \left\{ \max_{1 \leqslant i \leqslant n} \lambda_i(A) - \lambda_0,\ \lambda_0 - \min_{1 \leqslant i \leqslant n} \lambda_i(A) \right\}. \tag{17}$$

Let us prove that

$$\rho_0 = l\sqrt{1 - \bar{\lambda}(B)}, \tag{18}$$

where $\bar{\lambda}(B)$ is the maximum-modulus eigenvalue of the matrix

$$B = E - \frac{1}{l^2}(A - \lambda_0 E)^2. \tag{19}$$

According to (12), (19), and (17), we have

$$\lambda_i(B) = 1 - \frac{1}{l^2}(\lambda_i(A) - \lambda_0)^2 \geqslant 0 \ \forall\, i.$$

Hence

$$\bar{\lambda}(B) = 1 - \frac{1}{l^2}(\lambda^*(A) - \lambda_0)^2 = 1 - \frac{\rho_0^2}{l^2},$$

where $\lambda^*(A)$ is the eigenvalue of matrix A nearest to λ_0, that is, (18) holds true.

Chapter 4

METHODS OF SOLVING NONLINEAR EQUATIONS AND SYSTEMS

Sec. 24. The Iteration Method

Let us take an equation in one unknown x:

$$x = \varphi(x), \tag{1}$$

where φ is a given function of x.

If no constraints are imposed on the function φ, then the following situations may occur: the equation (1) may have either one solution or a finite number of solutions (more than one) or infinitely many solutions or, finally, the equation (1) may have no solution at all. Usually, we are interested at once in two points: (1) the existence of solutions and (2) how to find them (if any).

Formulated and proved below is a theorem which yields the sufficient conditions for the existence of a unique solution of the equation (1) on some interval. This theorem also indicates the method of finding an approximate solution (called the *iteration method*) and provides the estimates of the error of the approximate solution.

Let us first give the following definition.

Definition. On the interval $[a, b]$ the function φ is said to satisfy a *Lipschitz condition* (with constant α) if for any $x_1, x_2 \in [a, b]$ the following inequality is fulfilled:

$$|\varphi(x_1) - \varphi(x_2)| \leqslant \alpha |x_1 - x_2|. \tag{2}$$

Remark 1. In particular, if the function φ is continuously differentiable on the interval $[a, b]$, then on $[a, b]$ it satisfies a Lipschitz condition with the constant

$$\alpha = \max_{[a, b]} |\varphi'(x)|, \tag{3}$$

which follows readily from Lagrange's formula of finite increments.

Theorem 1. *Let on the interval* $[x_0, x_0 + r]$ *the function* φ *satisfy a Lipschitz condition with constant* α, *and*

$$0 < \alpha < 1, \tag{4}$$

$$0 \leqslant \varphi(x_0) - x_0 \leqslant (1 - \alpha)r. \tag{5}$$

Then on the interval $[x_0, x_0 + r]$ *the equation* (1) *has the unique solution*

$$x_* = \lim_{k \to \infty} x_k, \tag{6}$$

where x_0 *is the left-hand end point of the interval* $[x_0, x_0 + r]$,

$$x_k = \varphi(x_{k-1}), \quad k = 1, 2, \ldots . \tag{7}$$

Here, there are the following estimates:

$$|x_* - x_k| \leqslant \rho\alpha^k, \tag{8}$$

$$|x_* - x_k| \leqslant \frac{\alpha}{1 - \alpha} |x_k - x_{k-1}|, \tag{9}$$

where $\rho = \dfrac{\varphi(x_0) - x_0}{1 - \alpha} \leqslant r$, $k = 1, 2, \ldots$.

Proof. Let us first establish that the recurrent number sequence $\{x_k\}$ can in fact be found by the formula (7) and that this sequence is entirely arranged on the interval $[x_0, x_0 + \rho]$, which belongs to the given interval $[x_0, x_0 + r]$.

Suppose for simplicity that $x_0 = 0$. Then the interval $[x_0, x_0 + \rho]$ coincides with the interval $[0, \rho]$. With the choice of ρ taken into account the following equality will be fulfilled:

$$\varphi(0) = (1 - \alpha)\rho. \tag{10}$$

In the inequality (2) which, by hypothesis, is satisfied by the function φ on the interval $[0, r]$ and, in particular, on the interval $[0, \rho]$ we set $x_1 = x$, $x_2 = 0$ and replace the inequality thus obtained by the following two inequalities for the function φ

$$\varphi(0) - \alpha x \leqslant \varphi(x) \leqslant \varphi(0) + \alpha x, \tag{11}$$

these inequalities being equivalent on the interval $[0, \rho]$.

From the equality (10) and the right-hand inequality (11) it follows that on the interval $[0, \rho]$ the function φ obeys the following constraint:

$$\varphi(x) \leqslant \rho. \tag{12}$$

Suppose that the terms of the sequence $\{x_k\}$ with numbers $k = 0, 1, \ldots, m - 1$ have already been found in terms of the formula (7) and satisfy the condition

$$0 \leqslant x_k \leqslant \rho. \tag{13}$$

For instance, for $m = 2$ our suppositions are automatically fulfilled, since the points $x_0 = 0$ and $x_1 = \varphi(x_0) = \varphi(0)$ which are the terms of the sequence under consideration, by virtue of (10), obey the inequalities (13).

Whenever $x_{m-1} \in [0, \rho]$ we can find the next term x_m by the formula (7), where $k = m$. Let us show that it also satisfies the condition (13). Consider two cases:

(1) $0 \leqslant x_{m-1} \leqslant \min\{\rho, \varphi(0)/\alpha\}$. Using the equality (7) for $k = m$ and also the left-hand inequality (11), we obtain

$$x_m = \varphi(x_{m-1}) \geqslant \varphi(0) - \alpha x_{m-1}$$

$$\geqslant \varphi(0) - \alpha \min\left\{\rho, \frac{\varphi(0)}{\alpha}\right\} \geqslant \varphi(0) - \alpha \frac{\varphi(0)}{\alpha} = 0. \quad (14)$$

(2) $\varphi(0)/\alpha < x_{m-1} \leqslant \rho$. Subtracting the equality (7) with $k = m - 1$ from the equality (7) with $k = m$ and taking into account the Lipschitz condition (2), we find

$$|x_m - x_{m-1}| = |\varphi(x_{m-1}) - \varphi(x_{m-2})| \leqslant \alpha |x_{m-1} - x_{m-2}|.$$

Further, we analogously obtain

$$|x_{m-1} - x_{m-2}| \leqslant \alpha |x_{m-2} - x_{m-3}|$$

and so forth. After a finite number of steps we arrive at the inequality

$$|x_m - x_{m-1}| \leqslant \alpha^{m-1} |x_1 - x_0|. \quad (15)$$

Hence, taking into consideration that $|x_1 - x_0| = \varphi(0)$ and $x_{m-1} > \varphi(0)/\alpha \geqslant \alpha^{m-1}\varphi(0)$ (since $0 < \alpha < 1$), we find

$$x_m \geqslant x_{m-1} - |x_m - x_{m-1}| > 0. \quad (16)$$

Except for the above-considered cases (1) and (2), no other cases may occur for x_{m-1} which satisfies the condition (13) for $k = m - 1$. Consequently, from (14) and (16) the left-hand inequality (13) follows for $k = m$. The right-hand inequality (13) for $k = m$ follows from the equality $x_m = \varphi(x_{m-1})$, where $x_{m-1} \in [0, \rho]$, and from the inequality (12).

From what was proved by induction it follows that for $x_0 = 0$ the infinite sequence $\{x_k\}$ can be really constructed by the formula (7) and all of its terms satisfy the condition (13). If $x_0 \neq 0$, then it is similarly proved that the sequence $\{x_k\}$ found by the formula (7) is entirely on the interval $[x_0, x_0 + \rho]$.

Let us prove now that the sequence $\{x_k\}$ is fundamental. Taking into account the inequality (15) which is obviously true for any natural m and

arbitrary x_0 and the inequality $|x_1 - x_0| = |\varphi(x_0) - x_0| < r$ which follows from the conditions (4) and (5), we find

$$|x_{n+p} - x_n| \leqslant \sum_{m=n+1}^{n+p} |x_m - x_{m-1}| \leqslant \sum_{m=n+1}^{n+p} \alpha^{m-1}|x_1 - x_0|$$

$$< r \sum_{m=n+1}^{n+p} \alpha^{m-1} = r\alpha^n \frac{1-\alpha^p}{1-\alpha} < \alpha^n \frac{r}{1-\alpha}, \quad (17)$$

where n and p are any natural numbers. Since $\alpha^n \to 0$ as $n \to \infty$, the relationships (17) show that the sequence $\{x_k\}$ is fundamental. Therefore the limit (6) exists, and since the sequence $\{x_k\}$ is on the interval $[x_0, x_0 + \rho]$, it follows that $x_* \in [x_0, x_0 + \rho]$.

The given function φ satisfies the Lipschitz condition (2) on the interval $[x_0, x_0 + \rho]$ which, in particular, means that the function φ is continuous on this interval. This enables us to pass to the limit as $k \to \infty$ in the equality (7); we pass to the limit on the right under the sign of the function φ. As a result, we obtain the equality

$$x_* = \varphi(x_*) \quad (18)$$

which means that $x_* \in [x_0, x_0 + \rho]$ is a solution of the equation (1).

The existence of a solution of the equation (1) on the interval $[x_0, x_0 + \rho]$ and, hence, on the interval $[x_0, x_0 + r]$ containing the interval $[x_0, x_0 + \rho]$ has been proved.

Let us assume that the point $x_{**} \in [x_0, x_0 + r]$ is also a solution of the equation (1), that is,

$$x_{**} = \varphi(x_{**}). \quad (19)$$

Subtracting the equality (19) from the equality (18) and bearing in mind the Lipschitz condition (2), we obtain

$$|x_* - x_{**}| = |\varphi(x_*) - \varphi(x_{**})| \leqslant \alpha|x_* - x_{**}|,$$

that is,

$$|x_* - x_{**}| \leqslant \alpha|x_* - x_{**}|.$$

This inequality is possible under the condition (4) only if $x_{**} = x_*$.

The uniqueness of the solution of the equation (1) on the interval $[x_0, x_0 + r]$ has been also proved.

It remains to establish the estimates (8) and (9) for the error $x_* - x_k$ of the approximate solution x_k of the equation (1). Using the equalities (18) and (7), and also the Lipschitz condition (2), we find

$$|x_* - x_k| = |\varphi(x_*) - \varphi(x_{k-1})| \leqslant \alpha|x_* - x_{k-1}|$$
$$= \alpha|\varphi(x_*) - \varphi(x_{k-2})| \leqslant \alpha^2|x_* - x_{k-2}| \leqslant \ldots \leqslant \alpha^k|x_* - x_0|. \tag{20}$$

Since $x_* \in [x_0, x_0 + \rho]$, it follows that $|x_* - x_0| \leqslant \rho$. Combining this with (20) we obtain the estimate (8).

From the equalities (7) and (18) there follows the equality

$$x_* - x_{k-1} = x_k - x_{k-1} + \varphi(x_*) - \varphi(x_{k-1})$$

from which, in terms of the Lipschitz condition (2), we get

$$|x_* - x_{k-1}| \leqslant |x_k - x_{k-1}| + \alpha|x_* - x_{k-1}|$$

or

$$|x_* - x_{k-1}| \leqslant \frac{1}{1-\alpha}|x_k - x_{k-1}|.$$

Hence, taking advantage of the first inequality in the chain of the inequalities (20), we come to the estimate (9).

Theorem 1 has been proved completely.

Example. Consider the equation

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right),$$

where a is a number, $1/2 \leqslant a < 1$. This equation is an equation of the form (1) with $\varphi(x) = (1/2)(x + a/x)$. It has the solution $x_* = \sqrt{a}$, which can be readily verified.

Let us try to apply Theorem 1. We set $x_0 = a$ and $r = 1 - a$, that is, we choose $[x_0, x_0 + r] = [a, 1]$. We have for $1/2 \leqslant a < 1$:

$$\max_{[a, 1]} |\varphi'(x)| = \max_{[a, 1]} \left|\frac{1}{2}\left(1 - \frac{a}{x^2}\right)\right| = \frac{1-a}{2a} \leqslant \frac{1}{2}.$$

Consequently, on the interval $[a, 1]$ the function φ satisfies the Lipschitz condition (2) with $\alpha = 1/2$. Thus, the condition (4) is fulfilled.

Let us verify the condition (5). We find

$$\varphi(x_0) - x_0 = \frac{1}{2}(a + 1) - a = \frac{1}{2}(1 - a) = (1 - \alpha)r.$$

The condition (5) is also fulfilled. Theorem 1 guarantees that on the interval $[a, 1]$ the equation in question has the unique solution, namely, the one determined by us by guess, that is, $x_* = \sqrt{a}$. In accordance with (7), for computing it we may apply the iteration method:

$$x_0 = a, \; x_k = \frac{1}{2}\left(x_{k-1} + \frac{a}{x_{k-1}}\right), \quad k = 1, 2, \ldots .$$

Using the inequality (9) which is more exact than (8), it is appropriate to estimate the error in the process of computations.

Remark 2. Another variant of Theorem 1 is also valid when the interval $[x_0, x_0 + r]$ is replaced by the interval $[x_0 - r, x_0]$, instead of the condition (5) there appears the condition

$$0 \leqslant x_0 - \varphi(x_0) \leqslant (1 - \alpha)r, \tag{5*}$$

and the quantity ρ has the form $\rho = \dfrac{x_0 - \varphi(x_0)}{1 - \alpha}$.

Remark 3. The iterations (7) have a geometrical interpretation. The solution x_* of the equation (1) is the abscissa of the point of intersection of the straight line $y = x$ and the curve $y = \varphi(x)$. Convergent iterations are shown in Figs. 12 and 13.

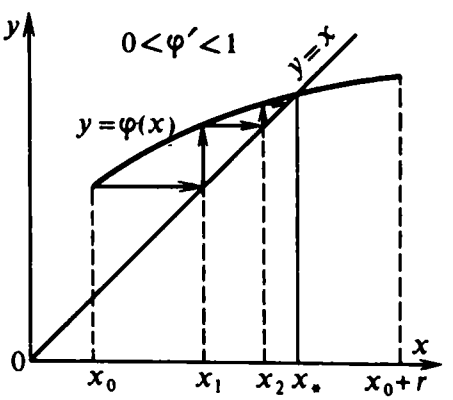


Fig. 12

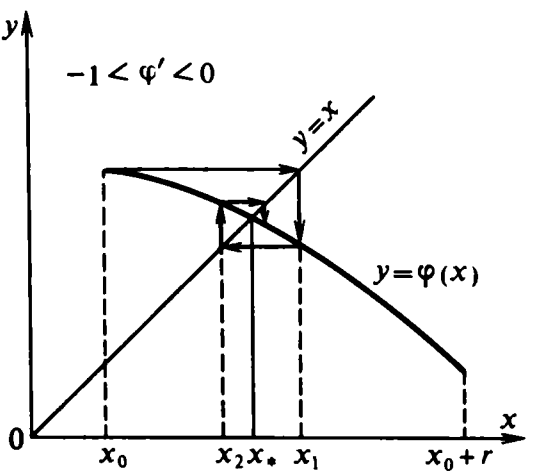


Fig. 13

Geometrically, it is seen that if in the neighbourhood of the solution x_* the inequalities $0 < \varphi'(x) \leqslant \alpha < 1$ are fulfilled, then the sequence $\{x_k\}$ converges monotonically to x_* from the side on which the initial approximation is located (see Fig. 12). In the case $-1 < -\alpha \leqslant \varphi'(x) < 0$, successive approximations are located in turn on different sides of the

solution x_* (see Fig. 13). In the last case it is possible to judge very simply the achieved accuracy by two successive approximations, namely, the deviation of x_k from x_* should not exceed $|x_k - x_{k-1}|$. It is also easily seen that with a decrease in $|\varphi'|$ the convergence becomes faster.

Remark 4. If the function φ entering into the equation (1) does not satisfy a Lipschitz condition with constant $\alpha < 1$, then the iterations (7) can diverge. For instance, let us consider the equation (1), where $\varphi(x) = bx$, $b > 1$, b is a number. Obviously, the function φ satisfies a Lipschitz condition with constant $\alpha = b > 1$ along the x-axis and does not satisfy this condition with any constant less than unity on any interval. The equation $x = bx$ has the only solution $x_* = 0$. But according to (7), for any $x_0 \neq 0$, $x_k = b^k x_0 \to \infty$ as $k \to \infty$.

Solving a System of Equations. In the n-dimensional vector space $\mathbf{R}^n$, let us introduce the distance*

$$\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \tag{21}$$

in terms of the norm (20.1) (see Item 8 of Introduction).

In accordance with (20.1) and (21), we have

$$\rho(\mathbf{x}, \mathbf{y}) = \begin{cases} \rho_1(\mathbf{x}, \mathbf{y}) = \max\limits_{1 \leqslant j \leqslant n} |x_j - y_j| \text{ or} \\ \rho_2(\mathbf{x}, \mathbf{y}) = \left(\sum\limits_{j=1}^{n} (x_j - y_j)^2 \right)^{1/2}. \end{cases} \tag{22}$$

The set $\bar{S}(\mathbf{y}^0; r) = \{\mathbf{x}: \rho(\mathbf{x}, \mathbf{y}^0) \leqslant r\}$ is called a *closed ball* of radius r with centre at the point $\mathbf{y}^0$ (here and elsewhere the indices in vectors will be given in the form of superscripts). Relying on Cauchy's criterion for sequences of numbers, we can prove that any closed ball $\bar{S}(\mathbf{y}^0; r)$, as well as the entire space $\mathbf{R}^n$, is a complete metric space with the metric (22).

Consider the equation (system of n, generally speaking, nonlinear equations in n unknowns)

$$\mathbf{x} = \boldsymbol{\varphi}(\mathbf{x}), \tag{23}$$

where $\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \ldots, \varphi_n(\mathbf{x}))$ is a given vector function of the variable $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ (variables $x_1, x_2, \ldots, x_n$), $\varphi_i(\mathbf{x}) = \varphi_i(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, n$. For instance, for $n = 2$ the system (23) in a more

---

* Prior to studying this subsection, the reader should be familiar with Sec. 1.5 of the book by Ya. S. Bugrov and S. M. Nikolsky *Differential Equations. Multiple Integrals. Series. Functions of a Complex Variable* (in English—Mir Publishers, Moscow, 1983).

detailed notation has the form

$$(x_1, x_2) = (\varphi_1(x_1, x_2), \varphi_2(x_1, x_2))$$

or

$$x_1 = \varphi_1(x_1, x_2), x_2 = \varphi_2(x_1, x_2),$$

where $\varphi_1(x_1, x_2)$ and $\varphi_2(x_1, x_2)$ are given functions of the variables x_1 and x_2.

The following theorem justifies the iteration method for solving the nonlinear system of equations (23).

Theorem 2. *Let a vector function $\varphi(\mathbf{x})$ be given in a closed ball $\bar{S} = \bar{S}(\mathbf{y}^0; r)$, and for any $\mathbf{x}, \mathbf{y} \in \bar{S}$ the following inequality be fulfilled:*

$$\rho(\varphi(\mathbf{x}), \varphi(\mathbf{y})) \leqslant \alpha\rho(\mathbf{x}, \mathbf{y}) \tag{24}$$

and, in addition,

$$\rho(\varphi(\mathbf{y}^0), \mathbf{y}^0) \leqslant (1 - \alpha)r, \tag{25}$$

where α is a number, $0 \leqslant \alpha < 1$. Then in $\bar{S}$ there is the unique solution $\mathbf{x}^*$ of the equation (23), and

$$\mathbf{x}^* = \lim_{k \to \infty} \mathbf{x}^k, \tag{26}$$

where $\mathbf{x}^0 \in \bar{S}$ is arbitrary,

$$\mathbf{x}^k = \varphi(\mathbf{x}^{k-1}), \quad k = 1, 2, \ldots. \tag{27}$$

In this case, the following inequalities are fulfilled:

$$\rho(\mathbf{x}^*, \mathbf{x}^k) \leqslant \alpha^k\rho(\mathbf{x}^*, \mathbf{x}^0) \leqslant 2\alpha^k r, \tag{28}$$

$$\rho(\mathbf{x}^*, \mathbf{x}^k) \leqslant \frac{\alpha}{1 - \alpha}\rho(\mathbf{x}^k, \mathbf{x}^{k-1}). \tag{29}$$

Proof. Let $\mathbf{x} \in \bar{S}(\mathbf{y}^0; r)$, that is, $\rho(\mathbf{x}, \mathbf{y}^0) \leqslant r$. Then applying the triangle inequality for distance, on the basis of (24) and (25), we get

$$\rho(\varphi(\mathbf{x}), \mathbf{y}^0) \leqslant \rho(\varphi(\mathbf{x}), \varphi(\mathbf{y}^0)) + \rho(\varphi(\mathbf{y}^0), \mathbf{y}^0)$$
$$\leqslant \alpha\rho(\mathbf{x}, \mathbf{y}^0) + (1 - \alpha)r \leqslant r.$$

Thus, if $\mathbf{x} \in \bar{S}(\mathbf{y}^0; r)$, then also $\varphi(\mathbf{x}) \in \bar{S}(\mathbf{y}^0; r)$, that is, the operator $\varphi(\mathbf{x})$ maps the complete metric space $\bar{S}(\mathbf{y}^0; r)$ into itself. Besides, by the condition (24), it is contracting. Consequently, by the contracting mapping principle, in $\bar{S}$ there is the only fixed point $\mathbf{x}^*$ of the operator $\varphi(\mathbf{x})$ which is the solution of the equation (system) (23). The relationships (26),

(28), and (29), except for the right-hand inequality in (28), are established in Sec. 1.5 of the above-mentioned book. The inequality $\alpha^k \rho(\mathbf{x}^*, \mathbf{x}^0) \leqslant 2\alpha^k r$ follows from the triangle inequality:

$$\rho(\mathbf{x}^*, \mathbf{x}^0) \leqslant \rho(\mathbf{x}^*, \mathbf{y}^0) + \rho(\mathbf{y}^0, \mathbf{x}^0) \leqslant r + r = 2r.$$

Theorem 2 has been proved.

The inequalities (28) imply that the iterations (26) converge to the desired solution $\mathbf{x}^*$ according to a geometric progression with denominator $\alpha < 1$. Since r and α are known, it is possible to predict sufficient number of iterations k for which the error $\rho(\mathbf{x}^*, \mathbf{x}^k)$ will be less than the assigned $\varepsilon > 0$. It is sufficient to require that the inequality $2\alpha^k r < \varepsilon$ or $\alpha^k < \varepsilon/2r$ be fulfilled. Hence $k > \ln(\varepsilon/2r)/\ln \alpha$, since $\ln \alpha < 0$. The minimum k is found by the formula

$$k = \max\left\{0, \left[\frac{1}{\ln \alpha} \ln \frac{\varepsilon}{2r}\right] + 1\right\}, \tag{30}$$

where $[a]$ is the integral part of a.

To obtain the estimate of the error $\rho(\mathbf{x}^*, \mathbf{x}^k)$, in the process of computations preference is given to the inequality (29) whose right-hand side is also expressed in terms of known quantities. This equality usually yields a more exact estimate than (28). This is due to the fact that the error for certain intermediate iterations may decrease more rapidly than α times. The estimate (28) will not correspond to this circumstance, while the right-hand side of (29) will automatically take it into account.

Remark 5. If the operator $\varphi(\mathbf{x})$ is contracting in the entire space $\mathbf{R}^n$, that is, if the condition (24) with $\alpha < 1$ is fulfilled for any $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$, then without the additional condition (25), from the contracting mapping principle there follow the existence of the solution of the system (23) and its uniqueness in $\mathbf{R}^n$. But even in this situation it is useful to impose the condition (25) which will be fulfilled if we assign some $\mathbf{y}^0$ (as close to the desired solution $\mathbf{x}^*$ as possible) and set $r = \rho(\varphi(\mathbf{y}^0), \mathbf{y}^0)/(1 - \alpha)$. Then, without solving the system (23), we may assert that in $\mathbf{R}^n$ the only solution $\mathbf{x}^*$ is really found in the closed ball $\bar{S}(\mathbf{y}^0; r)$, and it is possible to make use of the estimates (28) and (30).

Remark 6. The system of the linear algebraic equations (21.1) is a particular case of the system (23) with operator $\varphi(\mathbf{x}) = B\mathbf{x} + \mathbf{b}$, where B is a given matrix, and $\mathbf{b}$ is a given vector. If $\|B\| < 1$, then the given operator is contracting in the entire space $\mathbf{R}^n$. Indeed, let $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$. Then, by virtue of (21) and (20.8),

$$\rho(\varphi(\mathbf{x}), \varphi(\mathbf{y})) = \|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\| = \|B\mathbf{x} - B\mathbf{y}\|$$
$$= \|B(\mathbf{x} - \mathbf{y})\| \leqslant \|B\| \|\mathbf{x} - \mathbf{y}\| = \|B\| \rho(\mathbf{x}, \mathbf{y}),$$

that is, $\rho(\varphi(\mathbf{x}), \varphi(\mathbf{y})) \leqslant \alpha\rho(\mathbf{x}, \mathbf{y})$, where $\alpha = \|B\| < 1$. Consequently, for the successive approximations (21.2) the estimate (29) is valid which turns into the inequality (21.8) after replacing $\rho(\mathbf{x}^*, \mathbf{x}^k)$ by $\|\mathbf{x}^* - \mathbf{x}^k\|$ and $\rho(\mathbf{x}^k, \mathbf{x}^{k-1})$ by $\|\mathbf{x}^k - \mathbf{x}^{k-1}\|$.

Remark 7. For $n = 1$ Theorem 2 does not coincide with Theorem 1. The difference consists in the following. In Theorem 1 the equation (1) is considered either on the interval $[x_0, x_0 + r]$ or only on the interval $[x_0 - r, x_0]$ (see Remark 2) depending on the sign of the difference $\varphi(x_0) - x_0$. Here it is not required that the function φ be defined on the other interval. In Theorem 2 for $n = 1$ it is assumed that the function φ with the desired properties is specified on the interval $[x_0 - r, x_0 + r]$ (here y^0 is replaced by x_0).

Estimating the Magnitude of α. Suppose that the vector function $\varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \ldots, \varphi_n(\mathbf{x}))$ has continuous partial derivatives with respect to $x_1, x_2, \ldots, x_n$ in the closed ball $\bar{S} = \bar{S}(\mathbf{y}^0; r)$. Let us make the following notation:

$$\alpha_{ij} = \max_{\bar{S}} \left| \frac{\partial \varphi_i}{\partial x_j} \right|. \tag{31}$$

Let $\mathbf{x}, \mathbf{y} \in \bar{S}$. According to Lagrange's formula of finite increments, we have

$$\varphi_i(\mathbf{x}) - \varphi_i(\mathbf{y}) = \sum_{j=1}^{n} \frac{\partial \varphi_i(\xi^i)}{\partial x_j} (x_j - y_j), \tag{32}$$

$$i = 1, 2, \ldots, n,$$

where $\xi^i \in \bar{S}$ is some intermediate point.

With the aid of the matrix

$$\Phi = \begin{pmatrix} \dfrac{\partial \varphi_1(\xi^1)}{\partial x_1} & \cdots & \dfrac{\partial \varphi_1(\xi^1)}{\partial x_n} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial \varphi_n(\xi^n)}{\partial x_1} & \cdots & \dfrac{\partial \varphi_n(\xi^n)}{\partial x_n} \end{pmatrix} \tag{33}$$

the relationships (32) may be combined into one vector equality

$$\varphi(\mathbf{x}) - \varphi(\mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}),$$

where the matrix Φ on the right is multiplied by the vector $\mathbf{x} - \mathbf{y}$. Hence, in accordance with (21) and (20.8), we find

$$\rho(\varphi(\mathbf{x}), \varphi(\mathbf{y})) = \|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\| = \|\Phi(\mathbf{x} - \mathbf{y})\|$$

$$\leqslant \|\Phi\| \|\mathbf{x} - \mathbf{y}\| = \|\Phi\| \rho(\mathbf{x}, \mathbf{y}),$$

that is,

$$\rho_m(\varphi(\mathbf{x}), \varphi(\mathbf{y})) \leqslant \|\Phi\|_m \rho_m(\mathbf{x}, \mathbf{y}), \quad m = 1, 2,$$

where, depending on the choice of distance by the formula (22) using the first or the second method, the corresponding norm of the matrix (33) is taken.

Further, using the relationships (20.5) and (20.6), with (31) taken into account, we obtain

$$\rho_m(\varphi(\mathbf{x}), \varphi(\mathbf{y})) \leqslant \alpha_m \rho_m(\mathbf{x}, \mathbf{y}), \quad m = 1, 2, \tag{34}$$

where

$$\alpha_1 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} \alpha_{ij}, \quad \alpha_2 = \left(\sum_{i,j=1}^{n} \alpha_{ij}^2 \right)^{1/2}. \tag{35}$$

If for $m = 1$ or 2 it turns out that $\alpha_m < 1$, then with an appropriate choice of metric the condition (24) of Theorem 2 with $\alpha = \alpha_m$ will be fulfilled.

Example. It is required to find out whether a solution of the system of equations

$$\begin{aligned} x_1 &= \varphi_1(x_1, x_2) = 1.1 - \sin \frac{x_2}{3} + \ln \left(1 + \frac{x_1 + x_2}{5}\right), \\ x_2 &= \varphi_2(x_1, x_2) = 0.5 + \cos \frac{x_1 x_2}{6}. \end{aligned} \tag{36}$$

exists in the neighbourhood of the point $\mathbf{y}^0 = (y_1^0, y_2^0) = (1, 1)$.

Solution. We choose the metric $\rho_1(\mathbf{x}, \mathbf{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|\}$ and set $r = 1$. Then $\bar{S} = \bar{S}(\mathbf{y}^0; r) = \{\mathbf{x}: \rho_1(\mathbf{x}, \mathbf{y}^0) \leqslant 1\} = \{\mathbf{x}: 0 \leqslant x_i \leqslant 2, i = 1, 2\}$, that is, the closed ball $\bar{S}$ with a chosen concrete metric $\rho_1(\mathbf{x}, \mathbf{y})$ and $n = 2$ is a closed square in the usual sense.

We then find the partial derivatives

$$\frac{\partial \varphi_1}{\partial x_1} = \frac{1}{x_1 + x_2 + 5}, \quad \frac{\partial \varphi_1}{\partial x_2} = -\frac{1}{3} \cos \frac{x_2}{3} + \frac{1}{x_1 + x_2 + 5},$$

$$\frac{\partial \varphi_2}{\partial x_1} = -\frac{x_2}{6} \sin \frac{x_1 x_2}{6}, \quad \frac{\partial \varphi_2}{\partial x_2} = -\frac{x_1}{6} \sin \frac{x_1 x_2}{6}.$$

Hence, taking into account that $\sin(2/3) < 2/3$, we get

$$\max_{\bar{S}} \left|\frac{\partial \varphi_1}{\partial x_1}\right| = \frac{1}{5}, \quad \max_{\bar{S}} \left|\frac{\partial \varphi_1}{\partial x_2}\right| \leqslant \max\left\{\frac{1}{3} - \frac{1}{9}, \quad \frac{1}{5}\right\} = \frac{2}{9},$$

$$\max_{\bar{S}} \left|\frac{\partial \varphi_2}{\partial x_1}\right| < \frac{2}{6}\frac{2}{3} = \frac{2}{9}, \quad \max_{\bar{S}} \left|\frac{\partial \varphi_2}{\partial x_2}\right| < \frac{2}{9}.$$

In accordance with (34) and (35), the following magnitude may be taken as α:

$$\alpha = \max\left\{\frac{1}{5} + \frac{2}{9}, \quad \frac{2}{9} + \frac{2}{9}\right\} = \frac{4}{9}.$$

Thus, the condition (24) of Theorem 2 is fulfilled with $\alpha = 4/9 < 1$. Let us now find out whether the condition (25) is fulfilled. We have

$$\rho_1(\varphi(\mathbf{y}^0), \mathbf{y}^0) = \max\{|\varphi_1(1, 1) - 1|, \quad |\varphi_2(1, 1) - 1|\}. \tag{37}$$

Since $0 < \sin(1/3) < 1/3$, $0 < \ln(1 + 2/5) < 2/5$, $0 < \cos(1/6) < 1$, from (36) and (37) we obtain

$$\rho_1(\varphi(\mathbf{y}^0), \mathbf{y}^0) < 0.5 < (1 - \alpha)r = \frac{5}{9}.$$

The condition (25) is also fulfilled. By Theorem 2, the system (36) has the only solution x^* in the closed square $\bar{S}(\mathbf{y}^0; 1)$ being considered. The choice of $\mathbf{y}^0$ and r turned out to be successful.

The iterations (27) are carried out by the following formulas:

$$x_1^k = \varphi_1(x_1^{k-1}, x_2^{k-1}), \quad x_2^k = \varphi_2(x_1^{k-1}, x_2^{k-1}).$$

It is natural to take $\mathbf{x}^0 = \mathbf{y}^0 = (1, 1)$, i.e. $x_1^0 = x_2^0 = 1$ as the initial approximation. The accuracy of iterations, according to (29), is estimated by the inequality

$$\rho_1(\mathbf{x}^*, \mathbf{x}^k) \leqslant \frac{4}{5}\rho_1(\mathbf{x}^k, \mathbf{x}^{k-1})$$

$$= \frac{4}{5}\max\{|x_1^k - x_1^{k-1}|, |x_2^k - x_2^{k-1}|\}.$$

Remark 8. One of the conditions of the applicability of the iteration method is, roughly speaking, the smallness of the partial derivatives of the vector function $\varphi(\mathbf{x})$. In practice, it is required that at least one of the quantities of (35) be less than unity. But this is not sufficient for a solution to exist. There arises the problem of choosing a suitable $\mathbf{y}^0$ and r for which both conditions (24) and (25) of Theorem 2 are fulfilled. It is desirable to take the point $\mathbf{y}^0$ as close to the solution $\mathbf{x}^*$ as possible. The variation of r leads to that with an increase in r the quantities (31) or their upper bounds are usually increased and, consequently, α increases, but with a too small r the condition (25) may turn out to be impracticable. The problem of choosing $\mathbf{y}^0$ and r is a difficult one and should be treated individually in each case. A success is not guaranteed even if the system (23) has a solution, the more so since the existence of a solution may be unknown beforehand. And if there is no solution at all, then, naturally, it is impossible to choose the indicated $\mathbf{y}^0$ and r.

Sec. 25. Newton's Method

Consider the equation in one unknown x:

$$f(x) = 0. \tag{1}$$

Theorem 1. *If $f \in C_2[a, b]$, $f(a)f(b) < 0$, that is, if f attains values with opposite signs at the end points of the interval $[a, b]$ and f'' does not change sign on $[a, b]$, then the equation* (1) *has the only solution (root) x_* on $[a, b]$.*

The statement of the theorem is quite obvious. Under the conditions of Theorem 1 four cases are possible which are shown in Fig 14. The existence of the solution $x_* \in [a, b]$ follows from the continuity of f on $[a, b]$ and the supposition that $f(a)f(b) < 0$. The nonuniqueness of the solution under the condition $f(a)f(b) < 0$ would imply a change in sign of f''. This can be proved formally, but we shall confine ourselves mainly to geometrical considerations.

We take for x_0 the end point of the interval $[a, b]$ at which the function f has the same sign as f'' at those points of the interval $[a, b]$ where $f''(x) \neq 0$ (see Fig. 14). The case when $f''(x) \equiv 0$ is of little interest and therefore we are not going to dwell on it.

Let us denote by ω the interval whose end points are x_0 and x_*. Under the conditions of Theorem 1 the function f' is monotone on ω, and

$$m_1 = \min_{\omega} |f'(x)| = |f'(x_*)| > 0, \tag{2}$$

since otherwise the function f could not change sign as long as the sign of f'' is retained.

We set

$$M_1 = \max_{\omega} |f'(x)| = |f'(x_0)|, \tag{3}$$

$$M_2 = \max_{[a, b]} |f''(x)| > 0. \tag{4}$$

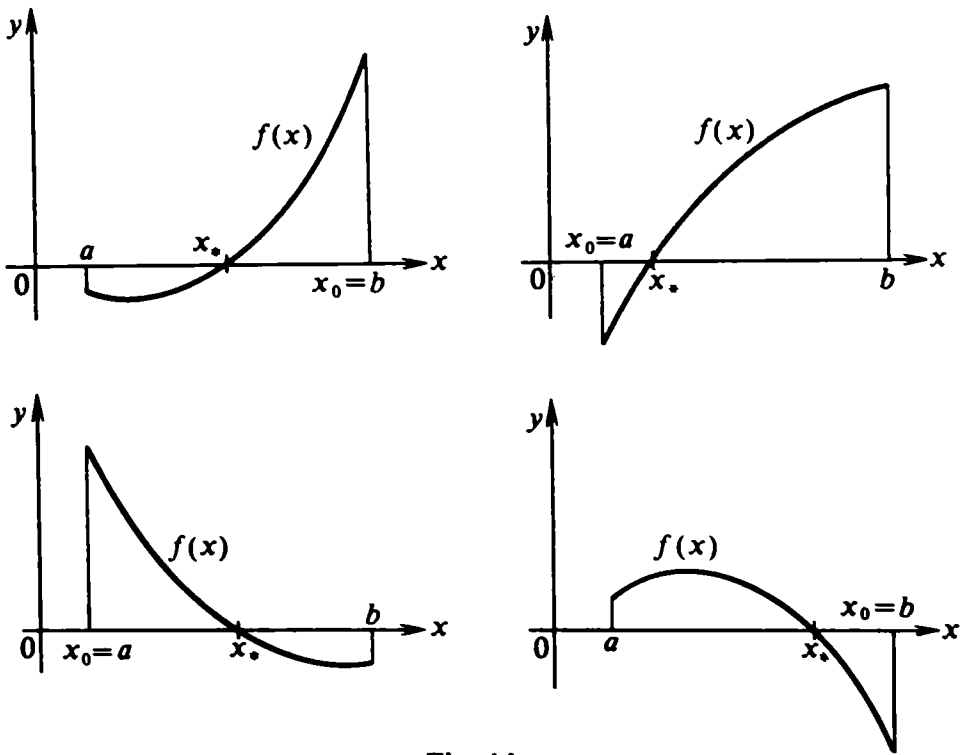


Fig. 14

Newton's Method. *Newton's method,* which is also called the *method of tangents,* consists in the following. Let us consider the tangent at the point x_0 to the curve $y = f(x)$ given by the equation

$$Y = f(x_0) + (x - x_0) f'(x_0).$$

Putting $Y = 0$, we find the point x_1 of intersection of the tangent with the axis of abscissas:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Constructing the tangent at the point x_1 (Fig. 15), we obtain by a similar formula the point x_2 of intersection of this tangent with the x-axis, etc.:

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}. \tag{5}$$

From geometrical considerations it is clear that under conditions of Theorem 1 the iterative sequence $\{x_k\}$ obtained by the formula (5) converges monotonically to the desired solution x_* of the equation (1).

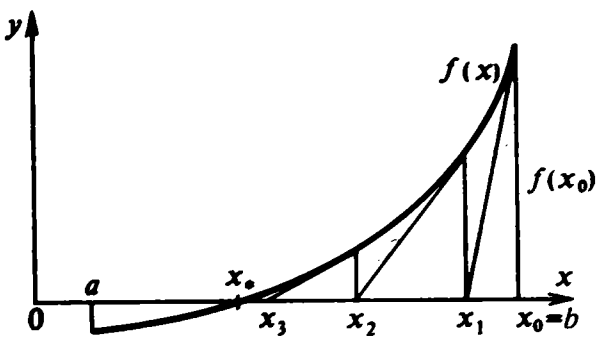


Fig. 15

Let us estimate the rate of convergence. Taking into account that $f(x_*) = 0$ and applying Lagrange's formula of finite increments, we find

$$|f(x_{k-1})| = |f(x_*) - f(x_{k-1})|$$

$$= |f'(\xi)|\,|x_* - x_{k-1}| \geqslant m_1 |x_* - x_{k-1}|,$$

where $\xi \in \omega$ is some point, and m_1 is the quantity (2). Combining this relationship with (3) and (5), we get

$$|x_k - x_{k-1}| = \frac{|f(x_{k-1})|}{|f'(x_{k-1})|} \geqslant \frac{m_1 |x_* - x_{k-1}|}{M_1}. \tag{6}$$

By virtue of the monotone convergence of the sequence $\{x_k\}$ to x_*, we have

$$|x_{k-1} - x_*| = |x_{k-1} - x_k| + |x_k - x_*|,$$

i.e.

$$|x_* - x_k| = |x_* - x_{k-1}| - |x_k - x_{k-1}|.$$

Hence, on the basis of (6), we come to the inequality

$$|x_* - x_k| \leqslant \alpha |x_* - x_{k-1}|, \tag{7}$$

where $0 \leqslant \alpha = 1 - m_1/M_1 < 1$.

The inequality (7) establishes that the error $x_* - x_k$ decreases at least according to a geometric progression with denominator $\alpha < 1$. This is characteristic for initial iterations. Then when the error becomes sufficiently small, the rate of convergence in Newton's method increases qualitatively. Let us convince ourselves of this. According to Taylor's formula, we have

$$0 = f(x_*) = f(x_{k-1}) + (x_* - x_{k-1}) f'(x_{k-1}) + \frac{(x_* - x_{k-1})^2}{2} f''(\xi),$$

where $\xi \in \omega$ is some point, that is,

$$x_* = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})} - (x_* - x_{k-1})^2 \frac{f''(\xi)}{2f'(x_{k-1})}. \tag{8}$$

Subtracting the equality (5) from the equality (8) and making use of the estimate $\left| \dfrac{f''(\xi)}{2f'(x_{k-1})} \right| \leqslant \beta$, where $\beta = M_2/(2m_1)$, and m_1, M_2 are the quantities (2) and (4), respectively, we arrive at the inequality

$$|x_* - x_k| \leqslant \beta (x_* - x_{k-1})^2.$$

Let us rewrite this inequality in the form

$$\beta |x_* - x_k| \leqslant (\beta (x_* - x_{k-1}))^2. \tag{9}$$

Hence it follows that as soon as the inequality $\beta |x_* - x_k| < 1$ is fulfilled for some k (by virtue of (7), this will happen without fail), the error multiplied by β will begin decreasing very rapidly according to the square law. After n next iterations we shall have

$$|x_* - x_{k+n}| \leqslant \frac{1}{\beta} (\beta (x_* - x_k))^{2^n}. \tag{10}$$

For instance, if $\beta |x_* - x_k| = 0.9$ and $n = 8$, then $2^n = 256$ and $(\beta (x_* - x_k))^{256} < 10^{-11}$.

Newton's Simplified Method. If the derivative $f'(x)$ is computed in a complicated way, then instead of the formula (5) the following formula is used:

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_0)}, \quad k = 1, 2, \ldots. \tag{11}$$

The convergence according to a geometric progression established by the inequality (7) is preserved here.

Relation Between the Iteration Method and Newton's Method. The equation (1) for $\lambda \neq 0$ is equivalent to the equation

$$x = x + \lambda f(x) \equiv \varphi(x).$$

The smaller $|\varphi'|$, the faster the iteration method converges. Let us require that $\varphi'(x_{k-1}) = 0$, i.e. $1 + \lambda f'(x_{k-1}) = 0$, and, consequently, $\lambda = -1/f'(x_{k-1})$. Then, in accordance with (24.7),

$$x_k = \varphi(x_{k-1}) = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})},$$

and we have arrived at Newton's method (5).

Solving a System of Equations. Let the vector function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))$, where $f_i(\mathbf{x}) = f_i(x_1, x_2, \ldots, x_n)$, be twice continuously differentiable in a neighbourhood of the solution $\mathbf{x}^*$ of the equation (a system of equations)

$$\mathbf{f}(\mathbf{x}) = 0. \tag{12}$$

The matrix $F(\mathbf{x})$ having the form

$$F(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \ldots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial f_n(\mathbf{x})}{\partial x_1} & \ldots & \dfrac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

is called the *Jacobian matrix* of the system of functions $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})$ at the point $\mathbf{x}$. If $\det F(\mathbf{x}) \neq 0$, let us denote by $F^{-1}(\mathbf{x})$ the inverse of the Jacobian matrix $F(\mathbf{x})$.

By Newton's method

$$\mathbf{x}^k = \mathbf{x}^{k-1} - F^{-1}(\mathbf{x}^{k-1}) f(\mathbf{x}^{k-1}), \quad k = 1, 2, \ldots. \tag{13}$$

By Newton's simplified method:

$$\mathbf{x}^k = \mathbf{x}^{k-1} - F^{-1}(\mathbf{x}^0) f(\mathbf{x}^{k-1}). \tag{14}$$

The simplification consists in that the inverse matrix $F^{-1}(\mathbf{x}^0)$ is found once but not in each iteration as in (13).

If $\det F(\mathbf{x}^*) \neq 0$ and the initial approximation $\mathbf{x}^0$ is taken sufficiently close to $\mathbf{x}^*$, then the iterations (13) and (14) converge in the metric (24.22) to $\mathbf{x}^*$. The nature of convergence is the same as for $n = 1$, that is, beginning with an instant, the iterations (13) converge very rapidly according to the square law, while for the iterations (14) the convergence is guaranteed only according to a geometric progression.

Sec. 26. The Method of Dividing a Line Segment into Two Equal Parts

Let $f \in C[a, b]$, $f(a)f(b) < 0$ and let it be known that the equation $f(x) = 0$ has the unique solution (unique root) $x_* \in [a, b]$. We set $a_0 = a$, $b_0 = b$, $c_0 = (a_0 + b_0)/2$, i.e. c_0 is the midpoint of the interval $[a_0, b_0]$, and compute $f(c_0)$. If $f(c_0) = 0$, then $x_* = c_0$ and no other computations are needed. If $f(c_0) \neq 0$, then the sign of $f(c_0)$ coincides either with the sign of $f(a_0)$ or with the sign of $f(b_0)$ as soon as $f(a_0)f(b_0) < 0$. Thus, at the end points of one of the two intervals $[a_0, c_0]$ or $[c_0, b_0]$ the function f has the same signs, and at the end points of the other opposite signs. We retain the interval at the end points of which f has opposite signs and reject the other interval since it does not contain the root x_*. We denote the retained interval by $[a_1, b_1]$, where

$$a_1 = \begin{cases} c_0, & \operatorname{sgn} f(a_0) = \operatorname{sgn} f(c_0), \\ a_0, & \operatorname{sgn} f(a_0) \neq \operatorname{sgn} f(c_0), \end{cases}$$

$$b_1 = \begin{cases} c_0, & \operatorname{sgn} f(b_0) = \operatorname{sgn} f(c_0), \\ b_0, & \operatorname{sgn} f(b_0) \neq \operatorname{sgn} f(c_0). \end{cases}$$

Obviously, $\operatorname{sgn} f(a_1) = \operatorname{sgn} f(a_0)$ and $\operatorname{sgn} f(b_1) = \operatorname{sgn} f(b_0)$. Therefore $f(a_1)f(b_1) < 0$. The desired root x_* is now on the interval $[a_1, b_1]$ which is of half length.

Further we proceed in a similar way. Suppose we have found some interval $[a_k, b_k] \subset [a, b]$ at the end points of which the function f has opposite signs and, consequently, which contains the desired root x_*. We find the midpoint of the interval $[a_k, b_k]$:

$$c_k = \frac{a_k + b_k}{2} \tag{1}$$

and compute $f(c_k)$. If $f(c_k) = 0$, then $x_* = c_k$. The computations have come to an end. If $f(c_k) \neq 0$, then we set

$$a_{k+1} = \begin{cases} c_k, & \operatorname{sgn} f(a_k) = \operatorname{sgn} f(c_k), \\ a_k, & \operatorname{sgn} f(a_k) \neq \operatorname{sgn} f(c_k), \end{cases}$$

$$b_{k+1} = \begin{cases} c_k, & \operatorname{sgn} f(b_k) = \operatorname{sgn} f(c_k), \\ b_k, & \operatorname{sgn} f(b_k) \neq \operatorname{sgn} f(c_k), \end{cases} \tag{2}$$

and so forth. This process may be finite if the midpoint of the interval obtained at some step coincides with the desired root x_*, otherwise this process is infinite. Figure 16 shows several initial steps. If computations are

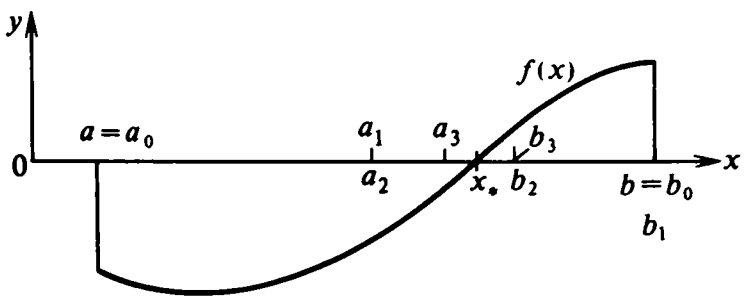


Fig. 16

brought to the kth step, then it is natural to take c_k as an approximate value for the desired root x_*. Here, the obvious error estimate is valid:

$$|x_* - c_k| \leqslant \frac{b-a}{2^{k+1}}. \tag{3}$$

The method discussed is a typical method used with computers, since the computations carried out by the formulas (1) and (2) turn out to be very simple and cyclic. It converges sufficiently rapidly. With each step the right-hand side of (3) decreases by half.

Sec. 27. The Method of Steepest (Gradient) Descent

Let $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. The system of n equations, $n \geqslant 2$,

$$\mathbf{f}(\mathbf{x}) = 0 \tag{1}$$

is equivalent to one equation

$$\Psi(\mathbf{x}) = 0, \tag{2}$$

where $\Psi(\mathbf{x}) = f_1^2(\mathbf{x}) + f_2^2(\mathbf{x}) + \ldots + f_n^2(\mathbf{x})$. Obviously, the points of the zero minima of the function $\Psi(\mathbf{x}) = \Psi(x_1, x_2, \ldots, x_n)$ are solutions of the equation (2). Let us assume that $\Psi(\mathbf{x})$ is twice continuously differentiable in the domain containing the isolated solution $\mathbf{x}^*$ in the neighbourhood of which the level surfaces of the function Ψ have the shape shown in Fig. 17.

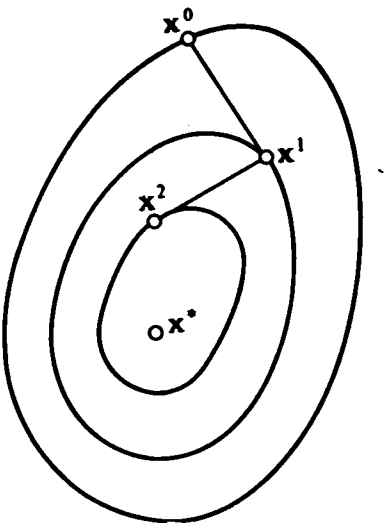


Fig. 17

Given the initial approximation $\mathbf{x}^0$, we seek the minimum of the function $\Psi(\mathbf{x}^0 - \lambda \operatorname{grad} \Psi(\mathbf{x}^0))$ in one variable λ. Actually, we find the minimum nonnegative root $\lambda = \lambda_0$ of the equation

$$\frac{d}{d\lambda} \Psi(\mathbf{x}^0 - \lambda \operatorname{grad} \Psi(\mathbf{x}^0)) = 0$$

using one of the above methods.

We set

$$\mathbf{x}^1 = \mathbf{x}^0 - \lambda_0 \operatorname{grad} \Psi(\mathbf{x}^0)$$

and, in general,

$$\mathbf{x}^k = \mathbf{x}^{k-1} - \lambda_{k-1} \operatorname{grad} \Psi(\mathbf{x}^{k-1}), \quad k = 1, 2, \ldots, \tag{3}$$

where λ_{k-1} is the minimum nonnegative root of the equation

$$\frac{d}{d\lambda} \Psi(\mathbf{x}^{k-1} - \lambda \operatorname{grad} \Psi(\mathbf{x}^{k-1})) = 0. \tag{4}$$

The convergence of the successive approximations (3) to the solution of the equation (2) is, generally speaking, not guaranteed, since we can get into a point of relative minimum.

Chapter 5

METHODS OF SOLVING THE BOUNDARY-VALUE PROBLEM FOR AN ORDINARY LINEAR SECOND-ORDER DIFFERENTIAL EQUATION

The present chapter contains two sections. Section 28 deals with the approximate methods of solving a boundary-value problem developed long before the appearance of computers and which are still of particular importance. Among them are the methods of collocation, of least squares, of subdomains, and also an effective and sufficiently universal Galerkin's method.

Section 29 gives a detailed description of the difference method of solving a boundary-value problem for an ordinary linear second-order differential equation. Despite the fact that this method is rather labour-consuming, it is widely recognized owing to its flexibility and universality. The difference method is rapidly developing thanks to the presence of highly efficient computers. Section 29 introduces the basic concepts of the theory of difference schemes, i.e. approximation, stability, convergence. In the same section the fundamental theorem is proved stating that convergence follows from approximation and stability. These concepts and theorem are illustrated by a difference scheme (a two-point boundary-value problem) for an ordinary differential equation.

Sec. 28. Methods of Minimizing the Discrepancy. Galerkin's Method

The *boundary-value problem* consists in the following. It is required to find the solution of the differential equation

$$Lu \equiv u'' + p(x)u' + q(x)u = f(x), \qquad a \leqslant x \leqslant b, \tag{1}$$

satisfying the two *boundary conditions:*

$$\begin{aligned} l_0 u &\equiv \alpha_0 u(a) + \beta_0 u'(a) = \gamma_0, \\ l_1 u &\equiv \alpha_1 u(b) + \beta_1 u'(b) = \gamma_1, \end{aligned} \tag{2}$$

where $p, q, f \in C[a, b]$ are given functions, $\alpha_j, \beta_j, \gamma_j$ are given numbers, and $\alpha_j^2 + \beta_j^2 > 0$, $j = 0, 1$.

In the general case, the boundary conditions (2) specify a linear relation between the value of the desired solution and its derivative at the end points of $[a, b]$ separately. In a particular case, if $\alpha_j = 1$ and $\beta_j = 0$, then the value of the desired solution is given at the appropriate end of the interval. Such a boundary condition is called the *condition of the first kind.* If $\alpha_j = 0$ and $\beta_j = 1$, then the value of the derivative of the solution is given at the end of the interval. This boundary condition is called the *condition of the second kind.* In the general case, when $\alpha_j \neq 0$ and $\beta_j \neq 0$, the boundary condition is called the *condition of the third kind.*

Unlike Cauchy's problem always having a unique solution, for the equation (1) in which the values of the solution and its derivative are given at one end point of the interval, the boundary-value problem (1), (2) may have either one solution or infinitely many solutions or, finally, may have no solution.

In order for the nonhomogeneous boundary-value problem (1), (2) *to have a unique solution, it is necessary and sufficient that the homogeneous boundary-value problem*

$$Lu \equiv u'' + p(x)u' + q(x)u = 0, \quad a \leqslant x \leqslant b, \tag{3}$$

$$\begin{aligned} l_0 u &\equiv \alpha_0 u(a) + \beta_0 u'(a) = 0, \\ l_1 u &\equiv \alpha_1 u(b) + \beta_1 u'(b) = 0 \end{aligned} \tag{4}$$

have only the trivial solution $u(x) \equiv 0$.

This condition cannot always be checked. In the next section we shall formulate simple sufficient conditions for the unique solvability of a boundary-value problem, and in the present section we shall assume the existence of a unique solution of a given boundary-value problem which often follows from physical considerations.

To find the approximate solution of the boundary-value problem (1), (2), we proceed in the following way. We assign on the interval $[a, b]$ a linearly independent system of twice continuously differentiable functions $\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots$ such that φ_0 satisfies the boundary conditions (2), that is, $l_0\varphi_0 = \gamma_0$, $l_1\varphi_0 = \gamma_1$, and the rest of the functions satisfy the homogeneous boundary conditions (4), that is, $l_0\varphi_i = 0, l_1\varphi_i = 0, i = 1, 2, \ldots$.

The assigned system of functions $\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots$ is termed the *base system.*

We compose a linear combination of $n + 1$ base functions

$$y_n(x) = \varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x) \tag{5}$$

with unknown (for the time being) coefficients $a_1, \ldots, a_n$. By virtue of the linearity of the operators l_0 and l_1, the function (5) satisfies the given boundary conditions (2) for any $a_1, \ldots, a_n$. Indeed,

$$l_j y_n = l_j \left(\varphi_0 + \sum_{i=1}^{n} a_i \varphi_i \right) = l_j \varphi_0 + \sum_{i=1}^{n} a_i l_j \varphi_i$$

$$= \gamma_j + 0 = \gamma_j, \quad j = 0, 1. \quad (6)$$

The function

$$\psi(x; a_1, \ldots, a_n) = L y_n(x) - f(x)$$

$$= L\varphi_0(x) - f(x) + \sum_{k=1}^{n} a_k L\varphi_k(x) \quad (7)$$

is called the *discrepancy*. The discrepancy is equal to the difference between the left-hand and right-hand sides of the equation (1) when on the left-hand side u is replaced by the function (5). As is seen from (7), the discrepancy is linearly dependent on the parameters $a_1, \ldots, a_n$ and is a characteristic of deviation of the function (5) from the unknown solution $u(x)$ of the boundary-value problem (1), (2). In any case, if for some values of the parameters $a_1, \ldots, a_n$ the discrepancy is identically equal to zero with respect to x on $[a, b]$, then the function (5) coincides with the solution of the boundary-value problem as soon as both the equation (1) and the boundary conditions (2) are satisfied.

But one usually fails to make the discrepancy identically equal to zero. Therefore one tries to choose the parameters $a_1, \ldots, a_n$ so as to make the discrepancy as small as possible. For the chosen $a_1, \ldots, a_n$, the function (5) is taken as the approximate solution of the boundary-value problem (1), (2).

There is a number of approximate methods of solving the boundary-value problem (1), (2) which differ in the technique for finding the parameters $a_1, \ldots, a_n$.

The Collocation Method. In the interval (a, b), n points $x_1, x_2, \ldots, x_n$ are fixed which are called the *collocation points* at which the discrepancy (7) is equated to zero:

$$\psi(x_1; a_1, \ldots, a_n) = 0,$$

$$\psi(x_2; a_1, \ldots, a_n) = 0,$$

$$\ldots\ldots\ldots\ldots\ldots$$

$$\psi(x_n; a_1, \ldots, a_n) = 0.$$

In a more detailed notation this system of linear algebraic equations with respect to $a_1, \ldots, a_n$ has the form

$$\begin{aligned} a_1 L\varphi_1(x_1) + \ldots + a_n L\varphi_n(x_1) &= f(x_1) - L\varphi_0(x_1), \\ a_1 L\varphi_1(x_2) + \ldots + a_n L\varphi_n(x_2) &= f(x_2) - L\varphi_0(x_2), \\ \ldots\ldots\ldots\

The Integral Method of Least Squares. The discrepancy is imposed with the requirement that the integral

$$I = \int_a^b \psi^2(x; a_1, \ldots, a_n)\, dx \tag{11}$$

take on the minimum value.

To attain the minimum of the integral, the following conditions should be fulfilled:

$$\frac{1}{2}\frac{\partial I}{\partial a_i} = \int_a^b \psi \frac{\partial \psi}{\partial a_i}\, dx = 0, \quad i = 1, 2, \ldots, n.$$

These conditions, with the expression (7) for ψ taken into account, lead to the following system of linear equations with respect to $a_1, \ldots, a_n$:

$$\begin{aligned}
&a_1(L\varphi_1, L\varphi_1) + \ldots + a_n(L\varphi_n, L\varphi_1) = (f - L\varphi_0, L\varphi_1),\\
&a_1(L\varphi_1, L\varphi_2) + \ldots + a_n(L\varphi_n, L\varphi_2) = (f - L\varphi_0, L\varphi_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_1(L\varphi_1, L\varphi_n) + \ldots + a_n(L\varphi_n, L\varphi_n) = (f - L\varphi_0, L\varphi_n),
\end{aligned} \tag{12}$$

where $(f, g) = \int_a^b f(x)\, g(x)\, dx$ is a scalar product.

If the system of functions $L\varphi_1, \ldots, L\varphi_n$ is linearly independent on the interval $[a, b]$, then, on the basis of Lemma 13.1, the system (12) has a unique solution.

The approximate solution of the boundary-value problem (9) found in the form (10) by the integral method of least squares has the expression

$$y_2(x) = 0.985(1 - x^2) - 0.078(x^2 - x^4).$$

The Discrete Method of Least Squares. Instead of the minimum of the integral (11), one seeks the minimum of the finite sum

$$\sum_{i=1}^{N} \psi^2(x_i; a_1, \ldots, a_n),$$

where $x_i \in (a, b)$ are some points, $N \geqslant n$. The system of equations for the coefficients of the approximate solution (5) has the form (12) with the on-

ly difference that the scalar product

$$(f, g) = \sum_{i=1}^{N} f(x_i)\, g(x_i)$$

be used. If $N = n$, then this method leads to the collocation method.

The approximate solution of the boundary-value problem (9) obtained in the form (10) by the discrete method of least squares for $N = 7$ and $x_i = -1 + i/4$ has the form

$$y_2(x) = 0.932(1 - x^2) - 0.047(x^2 - x^4).$$

The Method of Subdomains. Let $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$. The coefficients of the approximate solution (5) are found from the system of equations

$$\int_{x_{i-1}}^{x_i} \psi(x; a_1, \ldots, a_n)\, dx = 0, \quad i = 1, 2, \ldots, n. \tag{13}$$

The Galerkin Method. The method suggested by B.G. Galerkin is based on the requirement that the base functions $\varphi_1, \varphi_2, \ldots, \varphi_n$ be orthogonal to (7), that is,

$$\int_a^b \psi(x; a_1, \ldots, a_n)\, \varphi_i(x)\, dx = 0, \quad i = 1, 2, \ldots, n.$$

This requirement leads to the following system of linear algebraic equations for the coefficients of the approximate solution (5) of the boundary-value problem (1), (2):

$$\begin{aligned}
&a_1(L\varphi_1, \varphi_1) + \ldots + a_n(L\varphi_n, \varphi_1) = (f - L\varphi_0, \varphi_1),\\
&a_1(L\varphi_1, \varphi_2) + \ldots + a_n(L\varphi_n, \varphi_2) = (f - L\varphi_0, \varphi_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_1(L\varphi_1, \varphi_n) + \ldots + a_n(L\varphi_n, \varphi_n) = (f - L\varphi_0, \varphi_n),
\end{aligned} \tag{14}$$

where $(f, g) = \int_a^b f(x)\, g(x)\, dx$.

Example. Consider the boundary-value problem

$$Lu \equiv u'' + u = -x, \quad 0 \leqslant x \leqslant 1, \quad u(0) = 0, \quad u(1) = 0 \tag{15}$$

whose exact solution is $u(x) = \sin x/\sin 1 - x$.

Solution. We choose the base functions

$$\varphi_0(x) = 0,$$
$$\varphi_i(x) = x^i(1 - x), \quad i = 1, 2, \ldots, \tag{16}$$

which satisfy the given homogeneous boundary conditions. Let $n = 1$. Then the system (14) turns into one equation

$$a_1(\varphi_1'' + \varphi_1, \varphi_1) = (-x, \varphi_1),$$

that is,

$$a_1 \int_0^1 (-2 + x(1 - x))\, x(1 - x)\, dx = -\int_0^1 x^2(1 - x)\, dx.$$

The evaluation of the integrals yields $-(3/10)\, a_1 = -1/12$. Consequently, $a_1 = 5/18$, and the approximate solution has the form

$$y_1(x) = \frac{5}{18} x(1 - x).$$

Let us now take $n = 2$. The sought-for coefficients a_1 and a_2 satisfy the system of equations

$$a_1(L\varphi_1, \varphi_1) + a_2(L\varphi_2, \varphi_1) = (f, \varphi_1),$$
$$a_1(L\varphi_1, \varphi_2) + a_2(L\varphi_2, \varphi_2) = (f, \varphi_2)$$

which in our example takes the form

$$a_1 \int_0^1 (-2 + x(1 - x))\, x(1 - x)\, dx$$

$$+ a_2 \int_0^1 (2 - 6x + x^2(1 - x))\, x(1 - x)\, dx = -\int_0^1 x^2(1 - x)\, dx,$$

$$a_1 \int_0^1 (-2 + x(1 - x))\, x^2(1 - x)\, dx$$

$$+ a_2 \int_0^1 (2 - 6x + x^2(1 - x))\, x^2(1 - x)\, dx = -\int_0^1 x^3(1 - x)\, dx.$$

Evaluating the integrals, we find

$$\frac{3}{10}a_1 + \frac{3}{20}a_2 = \frac{1}{12}, \quad \frac{3}{20}a_1 + \frac{13}{105}a_2 = \frac{1}{20}.$$

Hence we obtain $a_1 = 71/369$ and $a_2 = 7/41$. The sought-for approximate solution has the form

$$y_2(x) = x(1 - x)\left(\frac{71}{369} + \frac{7}{41}x\right).$$

For the purpose of comparison, we give some values of the exact solution $u(x)$ of the boundary-value problem (15) with $y_1(x)$ and $y_2(x)$ being indicated:

x	u	y_1	y_2
0.25	0.044	0.052	0.044
0.50	0.070	0.069	0.069
0.75	0.060	0.052	0.060

Thus, the first approximation $y_1(x)$ has the error of order 0.01, and the second approximation $y_2(x)$ of order 0.001.

It is possible to prove that if the boundary-value problem

$$u'' + p(x)u' + q(x)u = f(x), \quad 0 \leqslant x \leqslant 1,$$
$$u(0) = 0, \quad u(1) = 0,$$

where $p, q, f \in C[0, 1]$, has the unique solution $u(x)$ (such as, for instance, the problem (15)), then making use of the base functions (16), the approximate solution $y_n(x)$ can be found by the Galerkin method for all sufficiently large n. This approximate solution converges uniformly to $u(x)$ on [0, 1] as $n \to \infty$. Moreover, $y_n'(x)$ converges uniformly to $u'(x)$ on [0, 1].

Sec. 29. The Difference Method. Basic Concepts of the Theory of Difference Schemes

Consider the boundary-value problem

$$Lu \equiv u'' + p(x)u' - q(x)u = f(x), \quad 0 \leqslant x \leqslant 1, \tag{1}$$
$$l_0u \equiv u(0) = \gamma_0, \quad l_1u \equiv u(1) = \gamma_1, \tag{2}$$

where $p, q, f \in C_2[0, 1]$ are given functions, $q(x) \geqslant 0$ on [0, 1], and γ_0, γ_1 are given numbers. This problem is a particular case of the boundary-value problem (28.1), (28.2).

Theorem 1. *The boundary-value problem* (1), (2) *has the unique solution* $u(x) \in C_4[0, 1]$.

The complete proof of Theorem 1 is beyond the scope of the present book. Let us only note that the indicated smoothness of the solution follows from the given smoothness of the functions p, q, and f. In the course of differential equations it is established that any solution of the equation (1) belongs to the class $C_4[0, 1]$, and hence the solution of the boundary-value problem (1), (2) also belongs to this class.

Let us set forth the difference method.

Grids and Grid Functions. Let us define on the interval [0, 1] a finite set of points (nodes): $\omega_h = \{x_j\}_{j=0}^{N}$, where $x_j = jh$, $h = 1/N$, and $N \geqslant 2$ is natural. The set ω_h is called the *grid*, and h is the *step* of the grid. The subset of the set ω_h obtained from ω_h by rejecting the extreme (boundary) *nodes* x_0 and x_N will be denoted by ω_h'. The set of ω_h' is also called the grid. The grid ω_h' is said to consist of *interior nodes* of the grid ω_h. Let also $\omega_h^* = \{x_0, x_N\}$ be the grid consisting of two *boundary nodes*. Thus, $\omega_h = \omega_h' \cup \omega_h^*$ (the grid ω_h is the union of the grids ω_h' and ω_h^*). The grids under consideration are shown in Fig. 18: the nodes belonging to the grid ω_h' are enclosed in small circles, and the boundary nodes forming the grid ω_h^* in small squares.

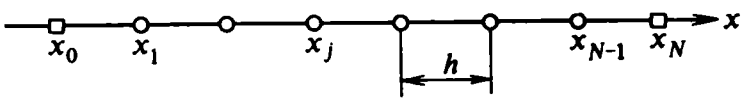


Fig. 18

The function y whose domain of definition is one of the grids ω_h, ω_h', or ω_h^* is called the *grid function*. The values of the grid function at nodal points will for brevity be denoted as follows: $y_j = y(x_j)$. The function f denoted throughout the interval [0, 1] generates a grid function attaining at the nodal point x_j the value equal to $f(x_j)$, $j = 0, 1, \ldots, N$. For the obtained grid function we retain the symbol f, and agree to denote its value at the nodal point x_j by f_j, $j = 0, 1, \ldots, N$.

Let us denote by Y_h, Y_h', and Y_h^* the sets of all the grid functions defined on ω_h, ω_h', and ω_h^*, respectively. The set Y_h (and also Y_h' and Y_h^*) is a linear space in which operations of multiplying a function by a number and of adding together functions are accomplished by usual rules. In the linear spaces Y_h, Y_h', and Y_h^* of grid functions we specify the respective norms

$$\|y\|_h = \max_{0 \leqslant j \leqslant N} |y_j|, \quad \|y\|_h' = \max_{1 \leqslant j \leqslant N-1} |y_j|, \quad \|y\|_h^* = \max\{|y_0|, |y_N|\}. \tag{3}$$

Difference Operators. The operator $L^h y$ is referred to as a *difference operator* if it associates each grid function $y \in Y_h$ with some grid function belonging to Y_h' (or Y_h^*).

Here are some examples of difference operators:

$$(1)\ (L^h y)_j = y_{j+1},\ j = 1, 2, \ldots, N-1;$$

$$(2)\ (l^h y)_j = y_j,\ j = 0, N;$$

$$(3)\ (L^h y)_j = \frac{y_{j-1} - 2y_j + y_{j+1}}{h^2},\ j = 1, 2, \ldots, N-1;\quad (4)$$

$$(4)\ (l^h y)_j = \begin{cases} \dfrac{y_1 - y_0}{h}, & j = 0, \\ \dfrac{y_N - y_{N-1}}{h}, & j = N. \end{cases}$$

The index j on the left-hand sides of the equalities indicates the number of the nodal point at which the operator attains the value given on the right. Whatever the grid function $y \in Y_h$, in Examples (1) and (3) we have $L^h y \in Y_h'$, and in Examples (2) and (4) we have $l^h y \in Y_h^*$.

For the sake of convenience, we gave the definition of a difference operator in a broad sense without requiring its explicit dependence on differences of first or higher orders, but only on the values of the grid function, also bearing in mind that the values of the grid function may be formally regarded as differences of zero order.

Let us consider the differential operator

$$Lu = u'' + p(x)u' - q(x)u \qquad (5)$$

which defines the left-hand side of the equation (1), and let us introduce a single boundary operator

$$lu = \begin{cases} l_0 u, & x = 0, \\ l_1 u, & x = 1, \end{cases} \qquad (6)$$

where, in accordance with the given boundary conditions (2), we shall regard for the time being that $l_0 u = u(0)$ and $l_1 u = u(1)$. Later on we shall consider a more general boundary operator.

In terms of the operators (5) and (6), the boundary-value problem (1), (2) can be written in the following compact form:

$$Lu = f, \qquad (7)$$

$$lu = \gamma, \qquad (8)$$

where

$$\gamma = \begin{cases} \gamma_0, & x = 0, \\ \gamma_1, & x = 1. \end{cases}$$

On the interval [0, 1] we have to approximate the given differential boundary-value problem (1), (2) (or, which is the same, the problem (7), (8)) by a difference boundary-value problem in which the grid function on the grid ω_h is sought for. Usually, we proceed as follows. First of all we construct some difference operators which approximate the differential operator Lu and boundary operator lu, and on the grid ω_h set a difference boundary-value problem usually called the *difference scheme* for the differential problem (1), (2). We then make sure that this difference scheme approximates the original boundary-value problem (1), (2), and investigate the stability of the difference scheme and the convergence of its solution to the solution of the differential problem as the step of the grid tends to zero. The difference scheme should also meet the following important requirement: for finding its solution, a sufficiently effective numerical method should be applied in practice. In what follows, the notions of approximation, stability, and convergence are defined more exactly.

We approximate the differential operator Lu on the grid ω_h' by the difference operator $L^h u$ obtained from the expression (5) by replacing at the nodal points of the grid ω_h' the derivatives u' and u'' by the appropriate difference derivatives according to the formulas (10.6) and (10.7). The operator $L^h u$ constructed in this way has the following form:

$$(L^h u)_j = \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} + p_j \frac{u_{j+1} - u_{j-1}}{2h} - q_j u_j, \quad (9)$$

$$j = 1, 2, \ldots, N - 1.$$

Let $u \in C_4[0, 1]$ be an arbitrary function, in particular, the solution of the boundary-value problem (1), (2). Then, on the basis of (10.2), (10.3), and (3), the following inequality is fulfilled:

$$\|Lu - L^h u\|_h' \leqslant c_u h^2, \quad (10)$$

where c_u is some constant independent of h.

Indeed, by virtue of (10.2) and (10.3), we have

$$\left| p(x_j)u'(x_j) - p_j \frac{u_{j+1} - u_{j-1}}{2h} \right| \leqslant \frac{h^2}{6} \|p\|_C \|u'''\|_C,$$

$$\left| u''(x_j) - \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} \right| \leqslant \frac{h^2}{12} \|u^{(4)}\|_C,$$

where $p_j = p(x_j)$, $j = 1, 2, \ldots, N-1$ and $\|p\|_C = \max\limits_{[0,\,1]} |p(x)|$. Combining this with (5) and (9), according to the definition (3) of the norm $\|\cdot\|_h'$, we readily obtain (10).

Definition. The difference operator $L^h u$ is said to *approximate* the differential operator Lu with kth order relative to the step h in the grid norm $\|\cdot\|_h'$ if $k > 0$ and for any sufficiently smooth function u the following inequality is fulfilled:

$$\|Lu - L^h u\|_h' \leqslant c_u h^k, \tag{11}$$

where c_u is some constant independent of h.

The foregoing difference operator $L^h u$ (9) approximates, in accordance with (10), the differential operator (5) with second order with respect to h.

The boundary operator (6) with the aid of which we specify the boundary conditions of the first kind (2), i.e. the conditions (8), is designed in a very simple way, namely, at the end points of the interval [0, 1] it takes on values coinciding with those of the function u. Therefore it is natural to approximate it by the difference operator (4) which attains the value at the nodes x_0 and x_N lying at the points of the interval [0, 1], these values coinciding with those of the grid function at these nodes.

Difference Schemes. Now, it is possible to set the following difference boundary-value problem (or more precisely, the family of difference boundary-value problems dependent on the parameter h):

$$L^h y = f, \tag{12}$$

$$l^h y = g, \tag{13}$$

where $L^h y$ is a difference operator defined by the formula (9), $l^h y$ is a difference operator (4), $f \in Y_h'$ is the grid function generated by the right-hand side of the equation (7), that is, the equation (1), and $g \in Y_h^*$ is the given grid function with values $g_0 = \gamma_0$ and $g_N = \gamma_1$.

The difference boundary-value problem (12), (13) is called the *difference scheme* for the differential boundary-value problem (7), (8). Taking into account the expressions (9) and (4) for the operators $L^h y$ and $l^h y$, we can write this difference scheme in a more detailed notation:

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + p_j \frac{y_{j+1} - y_{j-1}}{2h} - q_j y_j = f_j, \tag{14}$$

$$j = 1, 2, \ldots, N-1,$$

$$y_0 = \gamma_0, \quad y_N = \gamma_1. \tag{15}$$

It is readily seen that the difference boundary-value problem (14), (15) is a boundary-value problem of the form (22.1), (22.2) in which

$$A_j = \frac{1}{h^2} - \frac{p_j}{2h}, \quad B_j = \frac{1}{h^2} + \frac{p_j}{2h}, \quad C_j = \frac{2}{h^2} + q_j,$$

$$F_j = f_j, \quad \varkappa_0 = \varkappa_N = 0, \quad \nu_0 = \gamma_0, \quad \nu_N = \gamma_1.$$

Obviously, if

$$h < h_0 = \min\left\{1, \frac{2}{\|p\|_{C'}}\right\}, \tag{16}$$

then the conditions (22.3) are also fulfilled. Thus, if the step h of the grid ω_h satisfies the inequality (16), then, according to Lemma 22.1, the difference scheme (12), (13) has the unique solution $y \in Y_h$ which can be found by applying a very effective sweep method set forth in Sec. 22.

Approximation, Stability, and Convergence. The solution u of the differential problem (7), (8), generally speaking, does not satisfy the difference equations (12), (13) on the grid ω_h. But it is always possible to write the equalities

$$L^h u = f + \psi \text{ on } \omega_h', \tag{17}$$

$$l^h u = g + \varphi \text{ on } \omega_h^*, \tag{18}$$

where $\psi = L^h u - f$ and $\varphi = l^h u - g$ are grid functions called the *discrepancies* of the solution of the differential problem for the difference problem ($\psi \in Y_h'$ is the discrepancy for the difference equation (12) and $\varphi \in Y_h^*$ is the discrepancy for the boundary conditions (13)).

Definition. The difference scheme (12), (13) is said to *approximate* the differential problem (7), (8) on its solution u with kth order with respect to h if $k > 0$, and the norms of the discrepancies of the solution of the differential problem have the following magnitude:

$$\|\psi\|_h' = O(h^k), \|\varphi\|_h^* = O(h^k). \tag{19}$$

This definition is of a general nature, that is, the differential boundary-value problem (7), (8) may imply not necessarily the concrete boundary-value problem (1), (2), and not only the difference scheme (14), (15) may be taken as the difference scheme (12), (13). But we shall assume everywhere that the differential operators Lu and lu in the problem (7), (8) and the difference operators $L^h y$ and $l^h y$ in the problem (12), (13) are linear. Considered at the close of the present section is a boundary-value problem with more general boundary conditions than the above-mentioned ones with which the given definition goes very well.

Let us make sure that the difference scheme (14), (15) approximates the boundary-value problem (1), (2) on its solution u with second order with respect to h. Indeed, since $Lu = f$, we have

$$\psi = L^h u - f = L^h u - Lu \text{ on } \omega_h'$$

and, consequently, according to the inequality (10), where $L^h u$ is the operator (9), we have

$$\|\psi\|_h' = \|L^h u - Lu\|_h' = \|Lu - L^h u\|_h' = O(h^2).$$

Besides, in accordance with (18), (4), (13), and (2), we obtain

$$\|\varphi\|_h^* = \max\{|u_0 - g_0|, |u_N - g_N|\}$$
$$= \max\{|u(0) - \gamma_0|, |u(1) - \gamma_1|\} = 0 \ \forall h,$$

and the constant equal to zero is the quantity $O(h^2)$.

Remark 1. If the discrepancy φ equals zero on ω_h^* for any h, then the difference boundary conditions (13) approximate exactly the given boundary conditions (8) in the original problem.

The following important definition is also of a general nature.

Definition. The difference scheme (12), (13) is said to be *stable* if there is $h_0 > 0$ such that for any $h = 1/N < h_0$ and arbitrary grid functions $\xi \in Y_h'$ and $\eta \in Y_h^*$ the difference problem

$$L^h z = \xi, \tag{20}$$

$$l^h z = \eta \tag{21}$$

has the unique solution $z \in Y_h$, the following inequality being fulfilled

$$\|z\|_h \leqslant c'\|\xi\|_h' + c^*\|\eta\|_h^*, \tag{22}$$

where c' and c^* are certain constants independent both of h and of the functions ξ and η.

The stability (or instability) of a difference scheme is an inherent property of the difference scheme as such, that is, it is related in no ways to the original differential problem for which the difference scheme is constructed. It is of great importance that a difference scheme be stable. Such a difference scheme, in particular, should be stable with respect to round-off errors which often may be treated as additive components for the right-hand sides of the equations (12) and (13) in the form of certain grid functions $\xi \in Y_h'$ and $\eta \in Y_h^*$. By virtue of the linearity of the scheme, the error of the solution of the difference scheme (12), (13) coincides with the solution of the difference problem (20), (21) which, according to the inequality (22), will be small if the errors ξ and η or, strictly speaking, their

norms $\|\xi\|_h'$ and $\|\eta\|_h^*$ are small. It is also remarkable that the constants c' and c^* in the inequality (22) are independent of h, that is, we may say that the sensitivity of a stable difference scheme (12), (13) to the perturbations of the right-hand sides is not increased with grid refinement.

Examples of unstable difference schemes are given in Sec. 30. The following theorem testifies to the stability of the above-constructed difference scheme (14), (15).

Theorem 2. *The difference scheme* (12), (13), *where* $L^h y$ *is the difference operator defined by the formula* (9) *and* $l^h y$ *is the difference operator* (4), *is stable.*

Proof. It has been established above that the indicated difference scheme (12), (13), the more detailed notation of which has the form (14), (15), is uniquely solvable under the condition (16). Consequently, the difference boundary-value problem (20), (21) corresponding to the difference scheme (12), (13) is also uniquely solvable under the condition (16). Indeed, the difference scheme (12), (13) and the corresponding boundary-value problem (20), (21) are systems of linear algebraic equations differing only by constant terms. Therefore they are uniquely solvable simultaneously. It remains to prove the existence of the constants c' and c^* for which the solution of the problem (20), (21) satisfies the inequality (22).

By virtue of the linearity of the problem (20), (21), its solution z may be represented in the form

$$z = \lambda + \mu, \tag{23}$$

where λ and μ are the solutions of the difference problems

$$L^h\lambda = \xi, \quad l^h\lambda = 0, \tag{24}$$

$$L^h\mu = 0, \quad l^h\mu = \eta. \tag{25}$$

For the sake of simplicity, we will prove the inequality (22) only for the case when the function p is equal to zero everywhere. Then the difference problem (24), with (14), (15) in mind, may be rewritten as follows:

$$\lambda_{j-1} - (2 + h^2 q_j)\lambda_j + \lambda_{j+1} = h^2\xi_j, \quad j = 1, 2, \ldots, N-1,$$
$$\lambda_0 = 0, \quad \lambda_N = 0.$$

Since $q_j \geqslant 0, j = 1, 2, \ldots, N-1, Nh = 1$, on the basis of Lemma 22.2, for solving the problem (24) the following estimates are valid:

$$\|\lambda\|_h = \max_{0 \leqslant j \leqslant N} |\lambda_j| \leqslant N^2 \max_{1 \leqslant j \leqslant N-1} |h^2\xi_j|$$
$$= N^2h^2 \max_{1 \leqslant j \leqslant N-1} |\xi_j| = \max_{1 \leqslant j \leqslant N-1} |\xi_j| = \|\xi\|_h'. \tag{26}$$

Analogously, the problem (25) is written in the form

$$\mu_{j-1} - (2 + h^2 q_j)\mu_j + \mu_{j+1} = 0, \quad j = 1, 2, \ldots, N-1,$$

$$\mu_0 = \eta_0, \quad \mu_N = \eta_N.$$

According to Lemma 22.3, the norm for solving the problem (25) has the following expression:

$$\|\mu\|_h = \max_{0 \leqslant j \leqslant N} |\mu_j| = \max\{|\eta_0|, |\eta_N|\} = \|\eta\|_h^*. \tag{27}$$

Since, according to (23), $\|z\|_h \leqslant \|\lambda\|_h + \|\mu\|_h$, on the basis of (26) and (27), the solution of the problem (20), (21) for $p(x) \equiv 0$ satisfies the inequality (22) with the constants $c' = c^* = 1$. It is possible to prove that this inequality holds true for h which satisfies the inequality (16) also in the case of an arbitrary function $p \in C[0, 1]$, but with some other constants c' and c^*.

Let us give one more important definition.

Definition. When the grid is refined, the solution y of the difference scheme (12), (13) is said to *converge* to the solution u of the differential boundary-value problem (7), (8) with kth order with respect to h if $k > 0$ and

$$\|u - y\|_h = O(h^k). \tag{28}$$

In this case, we also say that the difference scheme has the kth *order of accuracy*.

Remark 2. This definition is also of a general nature, that is, as in the preceding case, the difference scheme (12), (13) implies not only the difference scheme (14), (15), and not only the problem (1), (2) may be taken as the differential problem (7), (8). Besides, the grid norm $\|\cdot\|_h$ used in the formula (28) may be defined in another way, that is, not in accordance with (3). But it should be borne in mind that, when the grid is refined, the presence of convergence in sense of (28) depends on the properties of the solution u of the original problem (7), (8), on the difference scheme (12), (13), and also on the choice of the norms in Y_h, Y_h', and Y_h^*.

Remark 3. When speaking of the convergence of the solution of a difference scheme, for the sake of brevity, we shall omit the words "when the grid is refined".

Now, it is possible to formulate the *principal theorem of the theory of difference schemes.*

Theorem 3. *Let the difference scheme* (12), (13) *approximate the differential boundary-value problem* (7), (8) *on its solution u with kth order with respect to h, and let it be stable. Then the solution of the difference*

scheme converges to the solution of the differential problem with the same order with respect to h.

Proof. Since, by the hypothesis, the difference scheme approximates the differential problem with kth order, there are, according to (19), numbers $h_1 > 0$ and $c_1 > 0$ such that for $h < h_1$ the following inequalities are fulfilled:

$$\|\psi\|_h' \leqslant c_1 h^k, \quad \|\varphi\|_h^* \leqslant c_1 h^k, \tag{29}$$

where ψ and φ are the discrepancies of the solution of the differential problem (see (17), (18)). Subtracting the equations (12) and (13) from the equalities (17) and (18), respectively, by virtue of the linearity of the operators L^h and l^h, we get

$$L^h(u - y) = \psi, \quad l^h(u - y) = \varphi. \tag{30}$$

Since, by the supposition, the difference scheme is stable, there are numbers h_0, $0 < h_0 \leqslant h_1$, c', and c^* such that for $h < h_0$ the difference problem (30) is uniquely solvable. On the basis of (22) and (29), the following inequalities are fulfilled:

$$\|u - y\|_h \leqslant c'\|\psi\|_h' + c^*\|\varphi\|_h^* \leqslant c'c_1 h^k + c^*c_1 h^k = ch^k,$$

where u is the solution of the differential boundary-value problem (7), (8), y is the solution of the difference scheme (12), (13), and $c = c_1(c' + c^*)$ is independent of h, $h < h_0$.

The principal theorem has been proved.

Remark 4. The definitions of approximation, stability, and convergence of the solution of a difference scheme and the above-proved principal theorem formulated briefly as "approximation plus stability is convergence" are of a general nature and will be a technique when considering difference schemes for other problems. The proposed way of investigating the difficult question of convergence by separating it into two independent parts (the verification of approximation and the investigation of stability) is generally accepted. In practice, it is usual to construct different variants of difference approximating schemes and to try to choose a stable difference scheme.

Remark 5. As far as the difference scheme (14), (15) is concerned, if it approximates the boundary-value problem (1), (2) with second order with respect to h (this has been established above) and is stable, according to Theorem 2, then, when the grid is refined, the solution y of the difference boundary-value problem (14), (15) converges to the solution u of the problem (1), (2) with second order with respect to h, that is,

$$\|u - y\|_h = O(h^2). \tag{31}$$

In other words, the difference scheme (14), (15) has the second order of accuracy. Note that, according to the definition of the norm $\|\cdot\|_h$ in (3), the left-hand side of the equality (31) is the maximum-modulus deviation on the grid ω_h of the solution y of the difference scheme (14), (15) from the solution u of the differential boundary-value problem (1), (2).

Approximating the Boundary Conditions of the Second and the Third Kind. Let us take the boundary conditions which are more general than (2):

$$\begin{aligned} l_0 u &\equiv u(0) - \beta_0 u'(0) = \gamma_0, \\ l_1 u &\equiv \alpha_1 u(1) + \beta_1 u'(1) = \gamma_1, \end{aligned} \tag{2*}$$

where α_1, β_i, and γ_i are given numbers, $\alpha_1 \geqslant 0$, $\beta_i \geqslant 0$, $i = 0, 1$, $\alpha_1 + \beta_1 > 0$. The boundary condition at the left end point of the interval [0, 1] may turn out to be a condition of the first kind if $\beta_0 = 0$ and of the third kind if $\beta_0 > 0$, and the boundary condition at the point $x = 1$ may also be a condition of the second kind ($\alpha_1 = 0$ and $\beta_1 = 1$).

With the same requirements to the functions p, q, and f as in the problem (1), (2), the boundary-value problem (1), (2*) also has the unique solution $u(x) \in C_4[0, 1]$. The boundary-value problem (1), (2) is a particular case of the problem (1), (2*) ($\alpha_1 = 1$ and $\beta_0 = \beta_1 = 0$).

To approximate the boundary conditions (2*), we proceed as follows. Let $u(x)$ be the solution of the boundary-value problem (1), (2*). Since

$$u_1 = u_0 + hu_0' + \frac{h^2}{2} u_0'' + O(h^3),$$

where $u_0^{(k)} = u^{(k)}(0)$,

$$u_0' = \frac{u_1 - u_0}{h} - \frac{h}{2} u_0'' + O(h^2) \tag{32}$$

or, more roughly,

$$u_0' = \frac{u_1 - u_0}{h} + O(h). \tag{33}$$

Taking into consideration that u satisfies the equation (1), we may write the following equality:

$$u_0'' = f_0 - p_0 u_0' + q_0 u_0.$$

Consequently, according to (33), we have

$$u_0'' = f_0 - p_0 \frac{u_1 - u_0}{h} + q_0 u_0 + O(h).$$

Substituting the expression found for u_0'' into (32) and then the expression obtained for u_0' into the first boundary condition of (2*), we get the following relationship:

$$u_0 - \beta_0 \left[\frac{u_1 - u_0}{h} + \frac{h}{2} \left(p_0 \frac{u_1 - u_0}{h} - q_0 u_0 \right) \right] = \gamma_0 - h \frac{\beta_0 f_0}{2} + O(h^2)$$

or, which is the same,

$$a_0 u_0 - b_0 u_1 = g_0 + O(h^2), \tag{34}$$

where

$$a_0 = 1 + b_0 + \beta_0 q_0 \frac{h}{2}, \quad b_0 = \beta_0 \left(\frac{1}{h} + \frac{p_0}{2} \right), \tag{35}$$

$$g_0 = \gamma_0 - h \frac{\beta_0 f_0}{2}. \tag{36}$$

Analogously, using the equation (1) and the second boundary condition of (2*), we find the following relation between the values u_N and u_{N-1} of the solution of the problem (1), (2) at the rightmost and last but one nodes of the grid ω_h:

$$a_1 u_N - b_1 u_{N-1} = g_N + O(h^2), \tag{37}$$

where

$$a_1 = \alpha_1 + b_1 + \beta_1 q_N \frac{h}{2}, \quad b_1 = \beta_1 \left(\frac{1}{h} - \frac{p_N}{2} \right), \tag{38}$$

$$g_N = \gamma_1 - h \frac{\beta_1 f_N}{2}. \tag{39}$$

Let us introduce the difference operator $l^h y$:

$$(l^h y)_j = \begin{cases} a_0 y_0 - b_0 y_1, & j = 0, \\ a_1 y_N - b_1 y_{N-1}, & j = N, \end{cases} \tag{40}$$

where the coefficients a_0, b_0, a_1, and b_1 are determined by the formulas (35) and (38).

Now, by the boundary-value problem (7), (8) we shall understand the boundary-value problem (1), (2*) regarding that the differential operator

Lu is determined as before by the formula (5) and the boundary operator lu is expressed by (6) in which the operators l_0u and l_1u are the differential operators specifying the left-hand sides of the boundary conditions (2*).

Let us associate the boundary-value problem (7), (8) with the difference scheme (12), (13), where the difference operator L^hy is defined as before by the formula (9), the operator l^hy is expressed by (40), and the grid function $g \in Y_h^*$ is given by the formulas (36) and (39).

The constructed difference scheme (12), (13) approximates the differential boundary-value problem (7), (8) on its solution u with second order with respect to h. Indeed, for solving the problem (7), (8) let us consider the equalities (17) and (18) with discrepancies ψ and φ. Since the difference equation (12) remains unchanged, we have, as it has been established before, $\|\psi\|_h' = O(h^2)$. The equality (18), with (40) taken into account, actually combines two equalities (34) and (37) from which it is clearly seen that $\varphi_0 = O(h^2)$, $\varphi_N = O(h^2)$ and, consequently, $\|\varphi\|_h^* = \max\{|\varphi_0|, |\varphi_N|\} = O(h^2)$.

In a more detailed notation, the difference scheme (12), (13) for the boundary-value problem (1), (2*) can be represented in the form of the difference equation (14) with the boundary conditions

$$a_0y_0 - b_0y_1 = g_0, \quad a_1y_N - b_1y_{N-1} = g_N, \tag{15*}$$

where the coefficients a_0, b_0, a_1, and b_1 as well as the constant terms g_0 and g_N are determined by the formulas (35), (38), (36), and (39).

It is possible to prove that the difference scheme under consideration ((12), (13)), that is, the difference boundary-value problem (14), (15*) is stable. We are not going to dwell on this in detail. Let us only note that if h satisfies the inequality (16), then the difference boundary-value problem (14), (15*) has a unique solution which can be found by the sweep method. This is guaranteed by the conditions (22.3) which for the coefficients of the difference equation (14) are, as it has been verified above, fulfilled. For the boundary conditions (15*) the requirements (22.3) for the indicated h are also fulfilled, since, according to (35) and (38), $a_0 > b_0 \geqslant 0$, $a_1 \geqslant b_1 \geqslant 0$, $a_1 + b_1 > 0$ and, consequently, the boundary conditions (15*) can be rewritten in the form (22.2), where $0 \leqslant \varkappa_0 = b_0/a_0 < 1$ and $0 \leqslant \varkappa_1 = b_1/a_1 \leqslant 1$.

Thus, if the difference scheme (14), (15*) approximates the boundary-value problem (1), (2*) with second order with respect to h and is stable, according to Theorem 3, then its solution y converges to the solution u of the boundary-value problem (1), (2*) with the rate $O(h^2)$, that is, the relationship (31) holds true.

Applying Runge's Rule. Suppose that $p, q, f \in C_4[0, 1]$, i.e. the functions p, q, and f entering into the equation (1) have an excess smoothness as compared with that for which we established the estimate (31) of the error of the difference solution of the boundary-value problem (1), (2*). Then for h satisfying the inequality (16) it is possible to obtain the following relationship on the grid ω_h:

$$u = y + v(x)h^2 + r, \tag{41}$$

where u is the solution of the differential boundary-value problem (1), (2*), y is the solution of the difference boundary-value problem (14), (15*), $v \in C_2[0, 1]$ is a function independent of h, and r is a grid function,

$$\|r\|_h = O(h^{2+m}), \tag{42}$$

with $m = 2$ for the boundary conditions of the first kind (2) and $m = 1$ for $\beta_0 + \beta_1 > 0$ in the boundary conditions (2*).

The existence of the relationship (41) enables us, as in the case of Cauchy's problem (see Sec. 18), to apply at the common nodes of the grids ω_h and $\omega_{h/2}$ Runge's rule for obtaining the approximate estimate of the error and to make the Richardson difference solution more precise. Besides, from the relationship (41) we may obtain the following formula:

$$u'(x_j) = \frac{y_{j+1} - y_{j-1}}{2h} + O(h^2) \tag{43}$$

which implies that with the aid of the difference solution y it is possible to find at the nodes of the grid ω_h' the derivative of the unknown solution of the boundary-value problem (1), (2*) to within $O(h^2)$.

Indeed, according to (41), we have

$$\frac{u_{j+1} - u_{j-1}}{2h} = \frac{y_{j+1} - y_{j-1}}{2h} + \frac{v_{j+1} - v_{j-1}}{2h} h^2 + \frac{r_{j+1} - r_{j-1}}{2h}. \tag{44}$$

Since $u \in C_3[0, 1]$, on the basis of (10.2), we obtain

$$\frac{u_{j+1} - u_{j-1}}{2h} = u'(x_j) + O(h^2). \tag{45}$$

Since $v \in C_1[0, 1]$, according to Lagrange's formula of finite increments, we have

$$\frac{v_{j+1} - v_{j-1}}{2h} = v'(\xi_j), \tag{46}$$

where $\xi_j \in (x_{j-1}, x_{j+1})$ is some point. Finally, by virtue of (42),

$$\frac{r_{j+1} - r_{j-1}}{2h} = O(h^2). \tag{47}$$

The formula (43) follows from (44)-(47).

Remark 6. From the equation (1) it is possible to find at the nodes of the grid ω_h' the second derivative of the solution of the boundary-value problem (1), (2*) to within $O(h^2)$ by substituting the approximate values u and u' obtained to within $O(h^2)$ into this equation.

Stability with Respect to the Right-hand Side. Along with the above definition of stability "on the whole", it is expedient to give one more definition.

Definition. The difference scheme (12), (13) is said to be *stable with respect to the right-hand side* if there are numbers $h_0 > 0$ and $c' > 0$ such that for any $h = 1/N < h_0$ and any grid function $\xi \in Y_h'$ the difference boundary-value problem

$$L^h z = \xi, \quad l^h z = 0 \tag{20*}$$

has the unique solution $z \in Y_h$, and

$$\|z\|_h \leqslant c' \|\xi\|_h'. \tag{22*}$$

The matter is that in many cases boundary conditions (and other additional conditions to be studied in the next chapter) are precisely approximated, that is, $\varphi = l^h u - g = 0$ (the discrepancy φ is equal to zero as the element Y_h^*). Then the error $u - y$ on the grid ω_h is the solution of the boundary-value problem

$$L^h(u - y) = \psi, \; l^h(u - y) = 0.$$

From the proof of Theorem 3 it is readily seen that for the convergence (with the equation (7) approximated by the difference equation (12)) it is sufficient that the difference scheme be stable only with respect to its right-hand side. For example, if $\beta_0 = \beta_1 = 0$ in the boundary conditions (2*), then these boundary conditions, being of the first kind, are precisely approximated in the difference scheme (14), (15*).

Chapter 6

DIFFERENCE SCHEMES FOR PARTIAL DIFFERENTIAL EQUATIONS

This chapter deals with difference schemes for linear equations with partial derivatives of first and second orders. The main subjects discussed here are: approximation and stability, and also algorithms for finding the solutions of difference schemes.

Sec. 30. Linear Equations with First-order Partial Derivatives

Statement of Cauchy's Problem. Find the function $u(x, t)$, continuous in the closed strip $-\infty < x < +\infty$, $0 \leqslant t \leqslant T$, which for $0 < t \leqslant T$ satisfies the differential equation

$$Lu \equiv \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, \tag{1}$$

and for $t = 0$, the initial condition

$$lu \equiv u(x, 0) = g(x), \tag{2}$$

where $a > 0$, $T > 0$ are numbers, g is a given function, twice continuously differentiable throughout the x-axis, and

$$\sup_{-\infty < x < +\infty} |g''(x)| = G_2 < +\infty. \tag{3}$$

The equation (1) is called a *linear differential equation with first-order partial derivatives.*

The exact solution of Cauchy's problem (1), (2) is the function

$$u(x, t) = g(x - at). \tag{4}$$

Indeed, $u_x'(x, t) = g'(x - at)$ and $u_t'(x, t) = -ag'(x - at)$. Consequently, the function (4) satisfies the equation (1). The fulfilment of the initial condition (2) is obvious.

Although the solution of Cauchy's problem (1), (2) is known and therefore approximate methods are not needed for it, but this problem is a suitable model with the aid of which we shall study difference schemes for partial differential equations.

Grids and Norms. Let us construct grids in the strip in which Cauchy's problem is posed (Fig. 19). Let $h > 0$ be the step with respect to x,

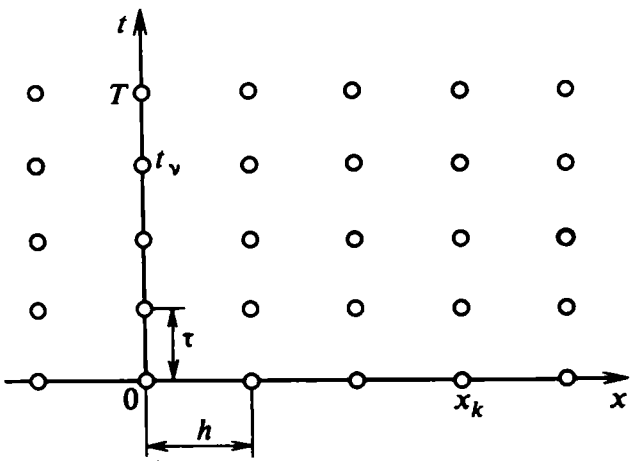


Fig. 19

$\tau = T/M$ (M natural) be the step with respect to time t, $x_k = kh$, $t_\nu = \nu\tau$, and $u_k^\nu = u(x_k, t_\nu)$. Each of the following three sets of points (nodes), i.e.

$$\omega_h = \{(x_k, t_\nu):\ k = 0, \pm 1, \ldots;\ \nu = 0, 1, \ldots, M\},$$
$$\omega_h' = \{(x_k, t_\nu):\ k = 0, \pm 1, \ldots;\ \nu = 1, 2, \ldots, M\},$$
$$\omega_h^* = \{(x_k, 0):\ k = 0, \pm 1, \ldots\},$$

is called the *grid*.

The grid ω_h^* consists of the nodes situated on the axis of abscissas where the initial condition (2) is specified. The grid ω_h is the union of the grids ω_h' and ω_h^*.

For the grid functions given on the grids ω_h and ω_h' let us introduce the respective norms

$$\|y\|_h = \sup_{\omega_h} |y_k^\nu|, \quad \|y\|_h' = \sup_{\omega_h'} |y_k^\nu|. \tag{5}$$

The Difference Scheme. Replacing the partial derivatives u_t' and u_x' in the equation (1) by difference derivatives (divided differences) with respect to t and x and substituting y for u, we arrive at the difference

scheme (at Cauchy's difference problem)

$$(L^h y)_k^\nu \equiv \frac{y_k^\nu - y_k^{\nu-1}}{\tau} + a\frac{y_k^{\nu-1} - y_{k-1}^{\nu-1}}{h} = 0, \tag{6}$$

$$(l^h y)_k \equiv y_k^0 = g_k. \tag{7}$$

Here and below, the index k traverses all integral values, $g_k = g(x_k)$, and $\nu = 1, 2, \ldots, M$.

The difference scheme (6), (7) can be written in brief notation:

$$L^h y = 0 \text{ on } \omega_h', \quad l^h y = g \text{ on } \omega_h^*. \tag{8}$$

Although the difference scheme depends on two parameters h and τ, for the sake of simplicity, we shall use only one index h for denoting the difference operators $L^h y$, $l^h y$, grids, and norms.

With k and ν fixed, the difference equation (6) contains only three unknowns $y_{k-1}^{\nu-1}$, $y_k^{\nu-1}$, and y_k^ν referring to the nodes $(x_{k-1}, t_{\nu-1})$, $(x_k, t_{\nu-1})$, and (x_k, t_ν), respectively. The locus of the points (nodes) at which the values of the function are used in the difference equation with k and ν fixed is called the *pattern* of the difference equation. The pattern of the difference equation (6) consisting of three nodes is shown in Fig. 20.

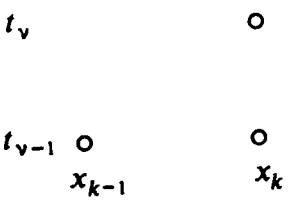


Fig. 20

Approximation. First of all let us note that the solution (4) of Cauchy's problem (1), (2) is twice continuously differentiable with respect to the variables x and t on the entire closed strip $-\infty < x < +\infty$, $0 \leqslant t \leqslant T$. Therefore, according to Taylor's formula, we have

$$u(x_k, t_\nu) = u(x_k, t_{\nu-1}) + \tau u_t'(x_k, t_{\nu-1}) + \frac{\tau^2}{2} u_{tt}''(x_k, \theta_{k\nu}),$$

$$u(x_{k-1}, t_{\nu-1}) = u(x_k, t_{\nu-1}) - h u_x'(x_k, t_{\nu-1}) + \frac{h^2}{2} u_{xx}''(\eta_{k\nu}, t_{\nu-1}),$$

that is,

$$\frac{u_k^\nu - u_k^{\nu-1}}{\tau} = u_t'(x_k, t_{\nu-1}) + \frac{\tau}{2} u_{tt}''(x_k, \theta_{k\nu}),$$

$$\frac{u_k^{\nu-1} - u_{k-1}^{\nu-1}}{h} = u_x'(x_k, t_{\nu-1}) - \frac{h}{2} u_{xx}''(\eta_{k\nu}, t_{\nu-1}),$$

where u is the solution of Cauchy's problem (1), (2), $t_{\nu-1} < \theta_{k\nu} < t_\nu$, and $x_{k-1} < \eta_{k\nu} < x_k$. Hence

$$(L^h u)_k^\nu \equiv \frac{u_k^\nu - u_k^{\nu-1}}{\tau} + a\frac{u_k^{\nu-1} - u_{k-1}^{\nu-1}}{h}$$

$$= u_t'(x_k, t_{\nu-1}) + a u_x'(x_k, t_{\nu-1}) + \frac{\tau u_{tt}''(x_k, \theta_{k\nu}) - ah u_{xx}''(\eta_{k\nu}, t_{\nu-1})}{2}. \quad (9)$$

Since the partial derivatives u_t' and u_x' of the solution of the problem (1), (2) are continuous on the closed strip $-\infty < x < +\infty, 0 \leqslant t \leqslant T$, by continuity this solution satisfies the equation (1) for $t = 0$ as well. Hence the equality

$$u_t'(x_k, t_{\nu-1}) + a u_x'(x_k, t_{\nu-1}) = 0 \quad (10)$$

is fulfilled for all k's and all ν's, for $\nu = 1$ included which corresponds to $t = 0$.

According to the definition (7) of the operator $l^h y$ and by virtue of (2), the solution u of the problem (1), (2) satisfies the initial condition on the grid ω_h^*

$$(l^h u)_k = u_k^0 = u(x_k, 0) = g(x_k) = g_k. \quad (11)$$

Thus, on the basis of (9)-(11), for the solution u of Cauchy's differential problem (1), (2) the following equalities are fulfilled on the grid $\omega_h = \omega_h' \cup \omega_h^*$:

$$(L^h u)_k^\nu = \psi_k^\nu \quad \text{on} \quad \omega_h', \quad (12)$$

$$(l^h u)_k = g_k \quad \text{on} \quad \omega_h^*, \quad (13)$$

where

$$\psi_k^\nu = \frac{\tau u_{tt}''(x_k, \theta_{k\nu}) - ah u_{xx}''(\eta_{k\nu}, t_{\nu-1})}{2} \quad (14)$$

is the discrepancy of the indicated solution for the difference equation (6). By virtue of (3)-(5) and (14), the discrepancy ψ satisfies the inequality

$$\|\psi\|_h' \leqslant c_0(h + \tau), \tag{15}$$

where $c_0 = a(1 + a)G_2/2$ is a constant independent of both h and τ. And as is seen from (7) and (13), the initial condition (2) is precisely approximated on ω_h^*. Thus, the difference scheme (6), (7) approximates Cauchy's problem (1), (2) on its solution (4) with first order with respect to h and τ.

The Computational Algorithm. Solving the equation (6) with respect to y_k^ν, we get

$$y_k^\nu = \left(1 - a\frac{\tau}{h}\right)y_k^{\nu-1} + a\frac{\tau}{h}y_{k-1}^{\nu-1}. \tag{16}$$

Since y_k^0 is known from the initial condition (7), using the formula (16), it is possible to compute "layer by layer" first y_k^1 of $k = 0, \pm 1, \ldots$, then y_k^2 for all k's, and so forth.

Stability. Consider the Cauchy difference problem

$$L^h z = \xi, \quad l^h z = 0, \tag{17}$$

where ξ is an arbitrary grid function defined on ω_h', and $\|\xi\|_h' < \infty$. Similarly to (16), the problem (17) can be written in the following form:

$$z_k^\nu = \left(1 - a\frac{\tau}{h}\right)z_k^{\nu-1} + a\frac{\tau}{h}z_{k-1}^{\nu-1} + \tau\xi_k^\nu, \tag{18}$$

$$z_k^0 = 0. \tag{19}$$

Suppose that

$$0 < a\frac{\tau}{h} \leqslant 1. \tag{20}$$

Then, obviously, $\left|1 - a\frac{\tau}{h}\right| + \left|a\frac{\tau}{h}\right| = 1$ and, consequently, in accordance with (18), we have

$$\sup_k |z_k^\nu| \leqslant \sup_k |z_k^{\nu-1}| + \tau\|\xi\|_h'. \tag{21}$$

Since, according to (19), $\sup_k |z_k^0| = 0$, from (21) we obtain $\sup_k |z_k^\nu| \leqslant \nu\tau\|\xi\|_h'$. Hence for the solution of the Cauchy difference problem (17)

we find the estimate

$$\|z\|_h = \sup_{\omega_h} |z_k^\nu| = \max_{0 \leqslant \nu \leqslant M} \sup_k |z_k^\nu|$$

$$\leqslant \max_{0 \leqslant \nu \leqslant M} \nu\tau \|\xi\|_h' = M\tau \|\xi\|_h' = T\|\xi\|_h'$$

where the constant T is independent of h, τ, and the grid function ξ. This estimate just means the stability of the difference scheme (8), that is, of the Cauchy difference problem (6), (7), by the right-hand side (provided that the condition (20) is fulfilled!).

Convergence. Since the initial condition (2) is precisely approximated on the grid ω_h^*, the difference equation (6), according to (12) and (15), approximates the differential equation (1) with first order with respect to h and τ, and, finally, the difference scheme (6), (7) is stable with respect to the right-hand side, then, by virtue of Theorem 29.3, with Remark 29.4 and the explanation to the theorem at the end of Section 29 taken into account, we have under the condition (20)

$$\|u - y\|_h = O(h + \tau),$$

where u is the solution of Cauchy's problem (1), (2), and y is the solution of the difference scheme (6), (7).

In other words, when the grid is refined, the solution of the difference scheme (6), (7) converges to the solution of Cauchy's problem (1), (2) with first order with respect to h and τ if $h \to 0$ and $\tau \to 0$ with the condition (20) observed.

Investigating Stability by the Perturbation Method. Let us show that the condition (20) is essential and that its violation may lead to the absence of stability. Suppose that $h \to 0$, $\tau \to 0$, and

$$a\frac{\tau}{h} = \text{const} = \lambda > 1. \tag{22}$$

In this case, obviously, approximation, that is, the inequality (15), is preserved, since when deriving the inequality (15), no constraints were imposed on the ratio of h and τ.

Consider Cauchy's difference problem (17) or, which is the same, the problem (18), (19), where

$$\xi_0^1 = \tau^m \varepsilon_0, \quad \xi_k^\nu = 0 \text{ for } (\nu - 1)^2 + k^2 \neq 0, \tag{23}$$

$\varepsilon_0 > 0$ and $m \geqslant 0$ are any fixed numbers. Thus, the function ξ is different from zero only at one point, that is, perturbation is given only at

one node $(0, \tau)$, and this perturbation is damping rather rapidly as $\tau \to 0$ if m is sufficiently large. According to (23),

$$\|\xi\|_h' = \sup_{\omega_h} |\xi_k^\nu| = |\xi_0^1| = \tau^m \varepsilon_0. \tag{24}$$

Let us set $\varepsilon = \tau^{m+1}\varepsilon_0$. Then, on the basis of (18), (19), (23), and (22), we shall have (Fig. 21)

$$z_0^1 = \tau\xi_0^1 = \tau^{m+1}\varepsilon_0 = \varepsilon, \qquad z_k^1 = 0 \text{ for } k > 0,$$

$$z_1^2 = \left(1 - a\frac{\tau}{h}\right)\cdot 0 + \lambda z_0^1 = \lambda\varepsilon, \qquad z_k^2 = 0 \text{ for } k > 1,$$

$$z_2^3 = \left(1 - a\frac{\tau}{h}\right)\cdot 0 + \lambda z_1^2 = \lambda^2\varepsilon, \qquad z_k^3 = 0 \text{ for } k > 2,$$

. . . . . .

$$z_{M-1}^M = \lambda^{M-1}\varepsilon = \lambda^{M-1}\tau^{m+1}\varepsilon_0 = \frac{\lambda^{M-1}}{M^{m+1}}T^{m+1}\varepsilon_0 \to \infty$$

for $\tau = 1/M \to 0$ $(M \to \infty)$, whatever the fixed $\varepsilon_0 > 0$ and $m \geqslant 0$ are.

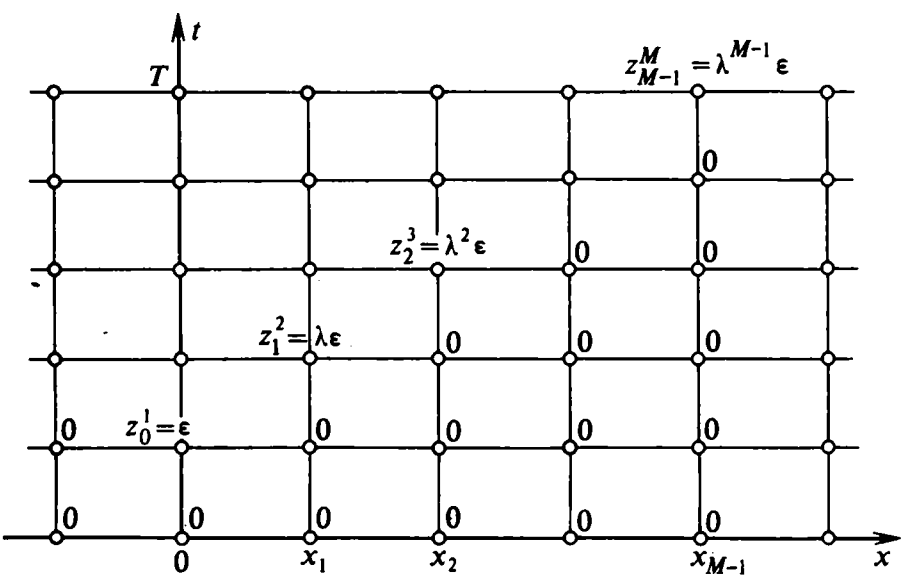


Fig. 21

Thus, according to (24), $\|\xi\|_h' = \tau^m\varepsilon_0$, that is, the norm of the function ξ is either constant if $m = 0$ or $\|\xi\|_h' \to 0$ for $\tau \to 0$ if $m > 0$. Nevertheless, $\|z\|_h \geqslant |z_{M-1}^M| \to \infty$ for $\tau \to 0$. Therefore the inequality of the form $\|z\|_h \leqslant c\|\xi\|_h'$ for solving the problem (17) which means the stability of the difference scheme (8) with respect to its right-hand side is

impossible for any constant c independent of h and τ. *Under the condition* (22) *the difference scheme* (6), (7) *is unstable, although the approximation remains.*

Unstable difference schemes are unfit for practical purposes because they are extremely sensitive to round-off errors. In the above-considered example, the function (23) multiplied by τ may be treated as the only round-off error due to computing y_0^1 by the formula (16), and z as the error of the difference solution which is the effect of the indicated round-off error. We saw that if the round-off error decreases at the rate τ^{m+1} as $\tau \to 0$, then for any m the error z occurring in the difference solution may increase under the condition (22) by the exponential law with respect to $M = T/\tau$, that is, very rapidly. It is not reasonable to demand that the round-off error decrease faster than by the power law with respect to τ, since the approximation error changes by the power law with respect to τ and h.

The Difference Scheme Unstable for Any Ratio of Steps. If in the equation (1) the derivative u_x' is replaced otherwise than when constructing the difference scheme (6), (7), namely, if we put

$$u_x' \approx \frac{u_{k+1}^{\nu-1} - u_k^{\nu-1}}{h}, \tag{25}$$

then, obviously, an approximation of order $O(h + \tau)$, that is, the inequality (15) is retained. But in this case, instead of the formula (16) the following formula will be obtained:

$$y_k^\nu = \left(1 + a\frac{\tau}{h}\right) y_k^{\nu-1} - a\frac{\tau}{h} y_{k+1}^{\nu-1}. \tag{26}$$

If $h \to 0, \tau \to 0$, and the ratio τ/h is constant, i.e. $\tau/h = \mu$, then since the first coefficient in the formula (26) exceeds unity ($\lambda = 1 + a\tau/h = 1 + a\mu > 1$) using the above-described perturbation method, the instability of the difference scheme under consideration is readily established for any μ.

Remark 1. At the first glance, the approximation of the derivative u_x' at the point $(x_k, t_{\nu-1})$ by the formula (25) is more natural than it is done in (6) with the aid of a backward divided difference. But, as we see, the more natural approximation may turn out to be less successful from the point of view of stability.

Sec. 31. Mixed Problem for the Heat Equation

Statement of a Mixed Problem. Let

$$\bar{D} = \{(x, t): 0 \leqslant x \leqslant 1, 0 \leqslant t \leqslant T\},$$

$$D' = \{(x, t): 0 < x < 1, 0 < t \leqslant T\},$$

that is, $\bar{D}$ is a closed rectangle, and D' is a half-open rectangle.

It is required to find the function $u(x, t)$ continuous on the closed rectangle $\bar{D}$ which on D' satisfies the heat equation

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \tag{1}$$

for $t = 0$ satisfies the initial condition

$$u(x, 0) = s(x), \tag{2}$$

and for $x = 0$ and $x = 1$ obeys the boundary conditions

$$u(0, t) = p(t), u(1, t) = q(t), \tag{3}$$

where $f(x, t), s(x), p(t)$, and $q(t)$ are given sufficiently smooth functions, and $s(0) = p(0), s(1) = q(0)$.

The problem (1)-(3) is said to be *mixed,* since it contains both the *initial condition* and *boundary conditions.* It is known that the posed problem has the unique solution $u(x, t)$. We shall assume that this solution has the partial derivatives $\partial u/\partial t$, $\partial^2 u/\partial t^2$, $\partial^2 u/\partial x^2$, and $\partial^4 u/\partial x^4$ on the closed rectangle $\bar{D}$.

Grids and Norms. Let $h = 1/N$ and $\tau = T/M$ be the steps with respect to x and t, where N, M are natural, $x_k = kh$, $t_\nu = \nu\tau$, and $u_k^\nu = u(x_k, t_\nu)$. Let us construct the grids (Fig. 22)

$$\omega_h = \{(x_k, t_\nu): k = 0, 1, \ldots, N, \nu = 0, 1, \ldots, M\},$$

$$\omega_h' = \{(x_k, t_\nu): k = 1, 2, \ldots, N - 1, \nu = 1, 2, \ldots, M\},$$

$$\omega_h^* = \omega_h \setminus \omega_h'.$$

The grid ω_h^* consists of the nodes of the grid ω_h marked on Fig. 22 by small crosses. These nodes are located on three sides of the rectangle $\bar{D}$ on which the initial and boundary conditions are given. The grid ω_h' consists of the remaining nodes of the grid ω_h. Let us take the following norms for the grid function defined on ω_h or on ω_h':

$$\|y\|_h = \max_{\omega_h} |y_k^\nu|, \quad \|y\|_h' = \max_{\omega_h'} |y_k^\nu|. \tag{4}$$

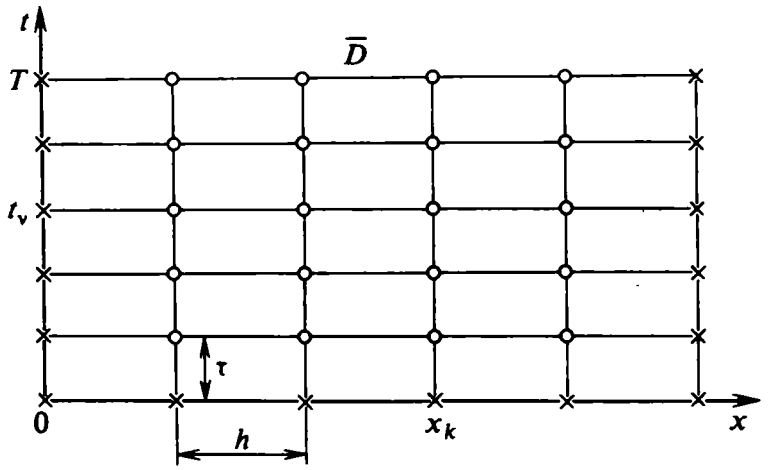


Fig. 22

Difference Schemes. Let us introduce the difference operator Λ:

$$\Lambda y_k^\nu = -\frac{y_{k-1}^\nu - 2y_k^\nu + y_{k+1}^\nu}{h^2}. \tag{5}$$

Here, the expression Λy_k^ν means the value of the grid function Λy at the point (x_k, t_ν), i.e. $(\Lambda y)_k^\nu$. To make the notation simpler, the parentheses are omitted. The analogous simplified notation will be used for introducing some other operators.

Let us take on the grid ω_h^* the identical operator

$$l^h y \equiv y \tag{6}$$

and the grid function

$$g = \begin{cases} s(x_k), & x = x_k, \quad t = 0, \quad k = 1, 2, \ldots, N-1, \\ p(t_\nu), & x = 0, \quad t = t_\nu, \quad \nu = 0, 1, \ldots, M, \\ q(t_\nu), & x = 1, \quad t = t_\nu, \quad \nu = 0, 1, \ldots, M. \end{cases} \tag{7}$$

Consider two difference schemes:

$$L_1^h y_k^\nu \equiv \frac{y_k^\nu - y_k^{\nu-1}}{\tau} + \Lambda y_k^{\nu-1} = f_k^{\nu-1}, \tag{8}$$

$$l^h y = g; \tag{9}$$

$$L_2^h y_k^\nu \equiv \frac{y_k^\nu - y_k^{\nu-1}}{\tau} + \Lambda y_k^\nu = f_k^\nu, \tag{10}$$

$$l^h y = g. \tag{11}$$

Here and henceforth, k varies from 1 to $N-1$ and $\nu = 1, 2, \ldots, M$.

The patterns of the difference equations (8) and (10) are shown in Figs. 23 and 24, respectively. Both difference schemes ((8), (9) and (10), (11)) are called *two-layer schemes,* since the patterns of the difference

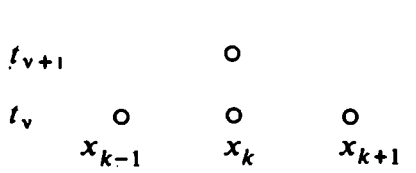


Fig. 23

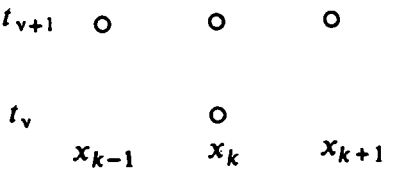


Fig. 24

equations (8) and (10) contain the nodes lying only on two *layers,* i.e. the subsets of the grid ω_h which correspond to the values of time $t = t_{\nu-1}$ and $t = t_\nu$. The layer found on the horizontal straight line $t = t_{\nu-1}$ is called the *lower layer,* and the other is called the *upper layer.* The difference schemes (8), (9) and (10), (11) differ in that in the equation (8) the operator Λ acts on the lower layer, while in the equation (10) it is taken out onto the upper layer and, besides, the values of the right-hand side $f_k^{\nu-1} = f(x_k, t_{\nu-1})$ and $f_k^\nu = f(x_k, t_\nu)$ are taken on different layers. For the time being, let us confine ourselves to the formal description of the two difference schemes. Their qualitative difference will be clarified below.

Approximation. Comparing the differential equation (1), on the one hand, and the difference equations (8) and (10), on the other, we see that to the partial derivative u_t' there corresponds the difference derivative $\dfrac{y_k^\nu - y_k^{\nu-1}}{\tau}$, and to the partial derivative $-u_{xx}''$ there corresponds a second-order difference dirivative in the direction x formed with opposite sign in terms of the operator Λ(5).

Let $u(x, t)$ be the solution of the problem (1)-(3). Since, by the assumption, its partial derivatives $\partial^2 u/\partial t^2$ and $\partial^4 u/\partial x^4$ are continuous and, consequently, are bounded on the closed rectangle $\bar{D}$, according to (5), (10.3), and (10.1), we have

$$\Lambda u_k^{\nu-1} = -u_{xx}''(x_k, t_{\nu-1}) + r_k^\nu, \tag{12}$$

$$\frac{u_k^\nu - u_k^{\nu-1}}{\tau} = u_t'(x_k, t_{\nu-1}) + \rho_k^\nu, \tag{13}$$

where $k = 1, 2, \ldots, M-1$ and $\nu = 1, 2, \ldots, N$,

$$|r_k^\nu| \leqslant c_1 h^2, \ |\rho_k^\nu| \leqslant c_2 \tau, \tag{14}$$

and c_1, c_2 are some constants independent of h, τ, k, and ν.

By virtue of the continuity of the partial derivatives u_t' and u_{xx}'' in $\bar{D}$, the solution of the problem (1)-(3) satisfies the equation (1) on the closed rectangle $\bar{D}$. Consequently, the equality

$$u_t'(x_k, t_{\nu-1}) - u_{xx}''(x_k, t_{\nu-1}) = f_k^{\nu-1} \tag{15}$$

is fulfilled for $k = 1, 2, \ldots, N-1$ and $\nu = 1, 2, \ldots, M$, that is, for $t_{\nu-1} = 0$, in particular.

According to (12), (13), and (15), the discrepancy ψ_1 of the solution u of the problem (1)-(3) for the difference equation (8) has the following form:

$$\psi_{1k}^{\nu} = L_1^h u_k^{\nu} - f_k^{\nu-1} = \frac{u_k^{\nu} - u_k^{\nu-1}}{\tau} + \Lambda u_k^{\nu-1} - f_k^{\nu-1}$$

$$= u_t'(x_k, t_{\nu-1}) + \rho_k^{\nu} - u_{xx}''(x_k, t_{\nu-1}) + r_k^{\nu} - f_k^{\nu-1} = r_k^{\nu} + \rho_k^{\nu}.$$

Hence, with (14) in mind, we obtain

$$\|\psi_1\|_h' = \max_{\omega_h'} |\psi_{1k}^{\nu}| = \max_{1 \leqslant \nu \leqslant M} \max_{1 \leqslant k \leqslant N-1} |r_k^{\nu} + \rho_k^{\nu}|$$

$$= O(h^2 + \tau). \tag{16}$$

In a similar way, we find

$$\|\psi_2\|_h' = O(h^2 + \tau), \tag{17}$$

where ψ_2 is the discrepancy of the solution u of the problem (1)-(3) for the difference equation (10).

Thus, both difference equations (8) and (10) approximate the differential equation (1) on the solution u of the problem (1)-(3) with second order with respect to h and first order with respect to τ.

The additional conditions, that is, the initial condition (2) and the boundary conditions (3), are precisely approximated on the grid ω_h^* with the aid of the identical operator l^h by the conditions (9) or, respectively, by the conditions (11), that is, the discrepancy of the solution u of the problem (1)-(3) for the conditions (9) as well as for (11) is equal to zero on the grid ω_h^*.

Thus, from the point of view of the approximation of the problem (1)-(3), both difference schemes (8), (9) and (10), (11) have the same guaranteed accuracy with respect to h and τ.

Computational Algorithms. Solving the difference equation (8) with respect to y_k^{ν}, we get

$$y_k^{\nu} = \frac{\tau}{h^2} y_{k-1}^{\nu-1} + \left(1 - \frac{2\tau}{h^2}\right) y_k^{\nu-1} + \frac{\tau}{h^2} y_{k+1}^{\nu-1} + \tau f_k^{\nu-1}. \tag{18}$$

Since y_k^0, y_0^ν, and y_N^ν, $k = 1, 2, \ldots, N-1$, $\nu = 0, 1, \ldots, M$, are known (they are specified on ω_h^* by the condition (9)), the solution of the difference scheme (8), (9) is found by the formula (18) explicitly, layer by layer. Therefore the difference scheme (8), (9) is said to be *explicit*.

The difference equation (10), with (5) taken into account, may be written in the form

$$\frac{\tau}{h^2} y_{k-1}^\nu - \left(1 + \frac{2\tau}{h^2}\right) y_k^\nu + \frac{\tau}{h^2} y_{k+1}^\nu = -y_k^{\nu-1} - \tau f_k^\nu. \quad (19)$$

According to (6), (7), and (11), we also have

$$y_0^\nu = p_\nu, \quad y_N^\nu = q_\nu. \quad (20)$$

Thus, if $y_k^{\nu-1}$, $k = 1, 2, \ldots, N-1$, are known (in particular, y_k^0, $k = 1, 2, \ldots, N-1$, are given by the condition (11)), then for finding the solution of the difference scheme (10), (11) on the next νth layer, it is necessary to solve the three-point difference equation (19) with the boundary conditions of the first kind (20), that is, a difference boundary-value problem of the form (22.1), (22.2). Therefore the difference scheme (10), (11) is called *implicit*.

To find the difference solution on the νth layer, we can apply the sweep method, since for the problem (19), (20) the conditions (22.3) are fulfilled (this should be checked by setting $k = j$, $y_k^\nu = z_j$, $y_{k\pm1}^\nu = z_{j\pm1}$, and $-y_k^{\nu-1} - \tau f_k^\nu = F_j$). In this case, according to (22.11), the number of arithmetic operations to be performed for finding the difference solution on one layer is equal to $O(N)$, that is, by the order with respect to N it is not greater than in applying an explicit formula (18) for the scheme (8), (9).

Stability and Convergence. Since the additional conditions (2) and (3) are precisely approximated in the difference schemes (8), (9) and (10), (11) on the grid ω_h^*, it suffices to investigate the stability only with respect to the right-hand sides. Let us first dwell on the difference scheme (8), (9). To investigate its stability with respect to the right-hand side, we have to consider the solution z of the auxiliary difference problem

$$L_1^h z_k^\nu \equiv \frac{z_k^\nu - z_k^{\nu-1}}{\tau} + \Lambda z_k^{\nu-1} = \xi_k^\nu, \quad (21)$$

$$l^h z = 0, \quad (22)$$

where ξ is an arbitrary grid function defined on ω_h'. Solving the difference

equation (21) with respect to z_k^ν, we obtain analogously to (18):

$$z_k^\nu = \frac{\tau}{h^2} z_{k-1}^{\nu-1} + \left(1 - \frac{2\tau}{h^2}\right) z_k^{\nu-1} + \frac{\tau}{h^2} z_{k+1}^{\nu-1} + \tau \xi_k^\nu, \quad (23)$$

$$k = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M.$$

Besides, in accordance with (22), we have

$$z_k^0 = 0, \quad k = 1, 2, \ldots, N-1;$$
$$z_0^\nu = z_N^\nu = 0, \quad \nu = 0, 1, \ldots, M. \quad (24)$$

Suppose that τ and h satisfy the following condition:

$$\frac{\tau}{h^2} \leqslant \frac{1}{2}. \quad (25)$$

Then, obviously,

$$\frac{\tau}{h^2} + \left|1 - \frac{2\tau}{h^2}\right| + \frac{\tau}{h^2} = 1.$$

The aforegoing and also (23) and (24) yield the inequality

$$\max_{0 \leqslant k \leqslant N} |z_k^\nu| \leqslant \max_{0 \leqslant k \leqslant N} |z_k^{\nu-1}| + \tau \max_{1 \leqslant k \leqslant N-1} |\xi_k^\nu|, \quad (26)$$

and since $\max\limits_{0 \leqslant k \leqslant N} |z_k^0| = 0$,

$$\max_{0 \leqslant k \leqslant N} |z_k^\nu| \leqslant \nu\tau \|\xi\|_h'.$$

Consequently,

$$\|z\|_h = \max_{0 \leqslant \nu \leqslant M} \max_{0 \leqslant k \leqslant N} |z_k^\nu| \leqslant M\tau \|\xi\|_h' = T\|\xi\|_h'$$

or, finally,

$$\|z\|_h \leqslant T\|\xi\|_h'. \quad (27)$$

The inequality (27) for solving the problem (21), (22) in which the constant T is independent of h, τ, and also of the function ξ means the stability of the difference scheme (8), (9) with respect to its right-hand side if the condition (25) is fulfilled. It is possible to prove that the violation of the condition (25) may lead to the violation of the stability of the difference scheme (8), (9). In particular, if $h \to 0$, $\tau \to 0$, and $\tau/h^2 \geqslant$ $\geqslant$ const $> 1/2$, then the difference scheme (8), (9) is unstable.

To investigate the stability of the difference scheme (10), (11), let us define an arbitrary grid function ξ on ω_h' and let us consider the difference problem

$$L_2^h z_k^\nu \equiv \frac{z_k^\nu - z_k^{\nu-1}}{\tau} + \Lambda z_k^\nu = \xi_k^\nu, \tag{28}$$

$$l^h z = 0, \tag{29}$$

without imposing any constraints on the ratio of the steps τ and h. Analogously to (19), (20), the problem (28), (29) can be written in the following form:

$$\frac{\tau}{h^2} z_{k-1}^\nu - \left(1 + \frac{2\tau}{h^2}\right) z_k^\nu + \frac{\tau}{h^2} z_{k-1}^\nu = -z_k^{\nu-1} - \tau \xi_k^\nu, \tag{30}$$

$$z_0^\nu = 0, \quad z_N^\nu = 0. \tag{31}$$

If $z_k^{\nu-1}$, $k = 1, 2, \ldots, N-1$, are known (in particular, by the condition (29), $z_k^0 = 0$, $k = 0, 1, \ldots, N$), then, as it has been noted above, the conditions (22.3) are fulfilled for the difference problem (30), (31), where ν is fixed. Consequently, by Lemma 22.1, this problem is uniquely solvable on the νth layer.

Obviously, there is k', $0 < k' < N$, such that

$$|z_{k'}^\nu| = \max_{0 \leqslant k \leqslant N} |z_k^\nu|. \tag{32}$$

Since $|z_{k'-1}^\nu| \leqslant |z_{k'}^\nu|$ and $|z_{k'+1}^\nu| \leqslant |z_{k'}^\nu|$, we have

$$|z_{k'}^\nu| \leqslant \left| z_{k'}^\nu \left(1 + \frac{2\tau}{h^2}\right) - \frac{\tau}{h^2}(z_{k'-1}^\nu + z_{k'+1}^\nu) \right|$$

and, consequently, in accordance with (30):

$$|z_{k'}^\nu| \leqslant |z_{k'}^{\nu-1}| + \tau |\xi_{k'}^\nu|.$$

From this inequality, with (32) taken into account, there follows the inequatity (26) and, in the final run, the estimate (27), which means that the difference scheme (10), (11) is stable with respect to its right-hand side for any ratio of the steps τ and h.

Thus, since the additional conditions (2) and (3) are precisely approximated on ω_h^*, then from the approximation (see (16), (17)) and the established stability with respect to the right-hand side, by virtue of the principal theorem of the theory of difference schemes (see Sec. 29), there follows the convergence of the solutions of the difference schemes (8), (9)

and (10), (11) to the solution of the problem (1)-(3) with second order with respect to h and first order with respect to τ, that is,

$$\|u - y\|_h = O(h^2 + \tau). \tag{33}$$

Here, in the case of the explicit scheme (8), (9) the constraint (25) is assumed to be fulfilled.

Definition. A difference scheme which is stable for any ratio of the steps τ and h is said to be *absolutely stable;* a difference scheme which is stable with the constraints imposed on τ and h is termed *conditionally stable.*

The shortage of the difference scheme (8), (9) is its conditional stability (the constraint (25) is rigid for the step τ with respect to time). Its advantage consists in the simplicity of computation by the explicit formula (18) and the possibility of being extended to Cauchy's problem (when the condition (2) is given throughout the x-axis and the boundary conditions (3) are absent). In the case of the mixed problem (1)-(3), preference is given to the implicit absolutely stable difference scheme (10), (11). When passing to each successive layer, the difference boundary-value problem (19), (20) is solved very effectively by the sweep method.

Sec. 32. The Wave Equation

Consider the mixed problem for the wave equation:

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \quad 0 < x < 1, \quad 0 < t \leqslant T, \tag{1}$$

$$u(x, 0) = p(x), \quad u_t'(x, 0) = q(x), \quad 0 \leqslant x \leqslant 1, \tag{2}$$

$$u(0, t) = 0, \quad u(1, t) = 0, \quad 0 \leqslant t \leqslant T, \tag{3}$$

where $f(x, t)$, $p(x)$, and $q(x)$ are given sufficiently smooth functions, and $p(0) = p(1) = q(0) = q(1) = 0$.

We shall assume that the problem (1)-(3) has the solution $u(x, t) \in C_4(\bar{D})$ and $\bar{D} = \{(x, t): 0 \leqslant x \leqslant 1, 0 \leqslant t \leqslant T\}$ is a closed rectangle. This solution is unique.

The Difference Scheme. We shall use the grids constructed on the closed rectangle $\bar{D}$ in Sec. 31, and the appropriate designation of the grid functions. Replacing approximately in the equation (1) the partial derivative u_{tt}'' by the second difference derivative in the direction of t, approximating the partial derivative $-u_{xx}''$ in terms of the difference

operator (31.5), and substituting y for u, we arrive at the difference equation

$$\frac{y_k^{\nu+1} - 2y_k^{\nu} + y_k^{\nu-1}}{\tau^2} + \Lambda y_k^{\nu} = f_k^{\nu}, \tag{4}$$

$$k = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M-1.$$

The pattern of the difference equation (4) is shown in Fig. 25. This equation can be solved explicitly with respect to $y_k^{\nu+1}$. But to find the

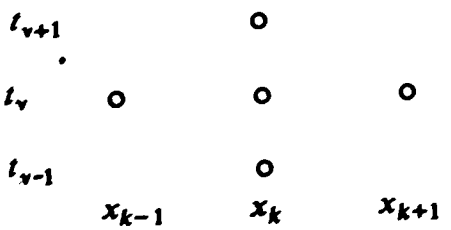


Fig. 25

values of the difference solution on the $(\nu + 1)$th layer, it is necessary to have the values of the desired solution having been computed on two previous layers. Therefore it is necessary to obtain the difference solution first separately on the layers corresponding to the values $\nu = 0$ and $\nu = 1$. In doing so, we shall be aided by the initial conditions (2).

First of all, using the first initial condition (2), we set

$$y_k^0 = p_k, \quad k = 1, 2, \ldots, N-1. \tag{5}$$

Besides, we put for $k = 1, 2, \ldots, N-1$

$$y_k^1 = p_k + \tau q_k + \frac{\tau^2}{2}(f_k^0 - \Lambda p_k). \tag{6}$$

The right-hand side of the formula (6) approximates Taylor's polynomial $u(x_k, 0) + \tau u_t'(x_k, 0) + \frac{\tau^2}{2} u_{tt}''(x_k, 0)$. Indeed, according to (2), $u(x_k, 0) = p_k$, $u_t'(x_k, 0) = q_k$, and from the equation (1) for the partial derivatives of the solution of the problem (1)-(3) there follows the relation $u_{tt}''(x_k, 0) = f(x_k, 0) + u_{xx}''(x_k, 0)$. The operator (31.5) is used on the right-hand side of this relation for approximating $-u_{xx}''(x_k, 0) = -p''(x_k)$.

Finally, according to the boundary conditions (3), we have

$$y_0^\nu = 0, \quad y_N^\nu = 0, \quad \nu = 0, 1, \ldots, M. \tag{7}$$

Now the difference scheme (4)-(7) is defined completely. This is an explicit three-layer scheme conditionally stable in some natural norms (see the pattern in Fig. 25).

If $h \to 0$, $\tau \to 0$, and $\tau/h \leqslant c < 1$, $c = \text{const}$, then the solution y of the difference scheme (4)-(7) converges to the solution u of the problem (1)-(3) in the following sense:

$$\|u - y\|_h = O(h^2 + \tau^2), \tag{8}$$

where $\|u - y\|_h = \max\limits_{0 \leqslant \nu \leqslant M} \left(h \sum\limits_{k=1}^{N-1} (u_k^\nu - y_k^\nu)^2 \right)^{1/2}$. The scheme (4)-(7) has second order with respect to both h and τ.

The Notion of the Method of Straight Lines. If in the problem (1)-(3) we introduce discreteness only with respect to x, then we arrive at the system of ordinary linear differential equations

$$\frac{d^2 y_k}{dt^2} - \frac{y_{k-1} - 2y_k + y_{k+1}}{h^2} = f(x_k, t), \tag{9}$$

$$k = 1, 2, \ldots, N - 1,$$

with the initial conditions

$$y_k(0) = p_k, \quad y_k'(0) = q_k, \tag{10}$$

and $y_0(t) \equiv y_N(t) \equiv 0$.

With our supposition concerning the smoothness of the solution u of the problem (1)-(3) taken into account, we have

$$\|u - y\|_h = O(h^2), \tag{11}$$

where $\|u - y\|_h = \max\limits_{0 \leqslant t \leqslant T} \left(h \sum\limits_{k=1}^{N-1} (u(x_k, t) - y_k(t))^2 \right)^{1/2}$ and $y_1(t), y_2(t), \ldots, y_{N-1}(t)$ is the solution of Cauchy's problem (9), (10).

This method is called the *method of straight lines*, since the approximate solution of the problem (1)-(3) is sought on the straight lines $x = x_k$, $k = 1, 2, \ldots, N - 1$, lying in the plane xt.

The difference method is often called the *grid* (or *net*) *method*.

Sec. 33. The Heat Equation in Two Spatial Variables

Let

$$\bar{R} = \{(x, y, t): 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1, 0 \leqslant t \leqslant T\},$$
$$R' = \{(x, y, t): 0 < x < 1, 0 < y < 1, 0 < t \leqslant T\},$$
$$S = \bar{R} \setminus R',$$

that is, $\bar{R}$ is a closed rectangular parallelepiped, R' is a half-open parallelepiped, S is the set of boundary points of the rectangular parallelepiped $\bar{R}$ which do not belong to R', and $f(x, y, t)$ is a sufficiently smooth function defined on $\bar{R}$.

It is required to find the function $u(x, y, t)$ continuous on $\bar{R}$ and satisfying on R' the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y, t) \tag{1}$$

and, in addition, obeying on S the additional condition

$$u = 0. \tag{2}$$

The condition (2) includes both the initial condition $u(x, y, 0) = 0$ for $t = 0$ and the homogeneous boundary conditions of the first kind for $x = 0$, $x = 1$, $y = 0$, and $y = 1$, that is, on the lateral faces of the parallelepiped $\bar{R}$.

The mixed problem (1), (2) has the unique solution $u(x, y, t)$.

Let us set $h = 1/N$, $\tau = T/M$, $x_k = kh$, $y_m = mh$, $t_\nu = \nu\tau$, and $f^\nu_{km} = f(x_k, y_m, t_\nu)$,

$$\Lambda_1 v^\nu_{km} = -\frac{v^\nu_{k-1,m} - 2v^\nu_{km} + v^\nu_{k+1,m}}{h^2}, \tag{3}$$

$$\Lambda_2 v^\nu_{km} = -\frac{v^\nu_{k,m-1} - 2v^\nu_{km} + v^\nu_{k,m+1}}{h^2}, \tag{4}$$

$$\Lambda v^\nu_{km} = \Lambda_1 v^\nu_{km} + \Lambda_2 v^\nu_{km}. \tag{5}$$

The Explicit Difference Scheme. The scheme

$$\frac{v^\nu_{km} - v^{\nu-1}_{km}}{\tau} + \Lambda v^{\nu-1}_{km} = f^{\nu-1}_{km}, \tag{6}$$

$$k, m = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M,$$

is stable for $\tau/h^2 \leqslant 1/4$ in uniform norms analogous to those which were introduced in Sec. 31 in the case of one spatial variable.

The Implicit Difference Scheme. The scheme

$$\frac{v^{\nu}_{km} - v^{\nu-1}_{km}}{\tau} + \Lambda v^{\nu}_{km} = f^{\nu}_{km}, \tag{7}$$

$$k, m = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M,$$

is absolutely stable, but on each layer with respect to time it is required to solve a system of $(N-1)^2$ equations.

In (6), (7) and below it is implied that the difference solution v at the points lying on S, in accordance with (2), is equal to zero.

The Method of Variable Directions or Fractional Steps. Both difference schemes (6) and (7) given above possess essential disadvantages, namely, one of them has a strict constraint on the step with respect to time τ, and when applying the other scheme, it is required to solve a system in $(N-1)^2$ unknowns on each layer with respect to time. Free of the indicated shortcomings is the following difference scheme, called the *scheme of variable directions* or *fractional steps:*

$$\frac{v^{\nu-1/2}_{km} - v^{\nu-1}_{km}}{\tau/2} + \Lambda_1 v^{\nu-1/2}_{km} + \Lambda_2 v^{\nu-1}_{km} = f^{\nu-1/2}_{km}, \tag{8}$$

$$\frac{v^{\nu}_{km} - v^{\nu-1/2}_{km}}{\tau/2} + \Lambda_1 v^{\nu-1/2}_{km} + \Lambda_2 v^{\nu}_{km} = f^{\nu-1/2}_{km}, \tag{9}$$

$$k, m = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M,$$

where $f^{\nu-1/2}_{km} = f(x_k, y_m, t_\nu - \tau/2)$.

In the difference scheme (8), (9) the step τ with respect to time is subdivided into two half-steps. The difference equation (8) corresponds to the first half-step, in it the quantities $v^{\nu-1}_{km}$ and $\Lambda_2 v^{\nu-1}_{km}$ are regarded as already known (in particular, $v^{0}_{km} = 0$, $k, m = 0, 1, \ldots, N$), and the unknowns have the superscript $\nu - 1/2$ (except for the right-hand side $f^{\nu-1/2}_{km}$ which is given). Multiplying the difference equation (8) by $-\tau/2$, we rewrite it in the following way:

$$\frac{\tau}{2h^2} v^{\nu-1/2}_{k-1,m} - \left(1 + \frac{\tau}{h^2}\right) v^{\nu-1/2}_{km} + \frac{\tau}{2h^2} v^{\nu-1/2}_{k+1,m} = F^{\nu-1/2}_{km}, \tag{10}$$

where $F^{\nu-1/2}_{km} = \frac{\tau}{2}(\Lambda_2 v^{\nu-1}_{km} - f^{\nu-1/2}_{km}) - v^{\nu-1}_{km}$ is known, and add the boundary conditions

$$v^{\nu-1/2}_{0m} = 0, \quad v^{\nu-1/2}_{Nm} = 0 \tag{11}$$

to the difference equation (10), in accordance with the condition (2).

We see that the difference problem (10), (11) falls into $N-1$ independent three-point difference boundary-value problems corresponding to each fixed m, $1 \leqslant m \leqslant N-1$. The difference boundary-value problem (10), (11) is solved by the sweep method (see. Sec. 22) for each m separately. The sweeping is realized with respect to the subscript k, that is, in the direction of the x-axis. According to (22.11), the solution of the difference problem (10), (11) requires $O(N)$ arithmetic operations for one value of m, and hence, $O(N^2)$ arithmetic operations are spent for solving $N-1$ problems.

After all the unknowns $v_{km}^{\nu-1/2}$ are found on the intermediate layer numbered $\nu-1/2$, we transpose them to the right in the difference equation (9) which corresponds to the second half-step. We then rewrite this difference equation in the form

$$\frac{\tau}{2h^2} v_{k,m-1}^{\nu} - \left(1 + \frac{\tau}{h^2}\right) v_{km}^{\nu} + \frac{\tau}{2h^2} v_{k,m+1}^{\nu} = F_{km}^{\nu}, \quad (12)$$

where $F_{km}^{\nu} = \frac{\tau}{2}(\Lambda_1 v_{km}^{\nu-1/2} - f_{km}^{\nu-1/2}) - v_{km}^{\nu-1/2}$ is known, and add the boundary conditions

$$v_{k0}^{\nu} = 0, \quad v_{kN}^{\nu} = 0 \quad (13)$$

to the difference equation (12), in accordance with the condition (2).

The problem (12), (13) also falls into $N-1$ three-point difference boundary-value problems corresponding to different fixed k, $1 \leqslant k \leqslant N-1$. Each of these problems is solved by the sweep method. Now the sweeping is performed with respect to the subscript m, that is, in the direction of the y-axis. The solution of the difference problems (12), (13) requires $O(N^2)$ arithmetic operations for all k's.

Thus, when passing from the $(\nu-1)$th layer to the νth layer according to the difference scheme (8), (9), the number of arithmetic operations to be performed is of the order of the number of sought-for unknowns on a single layer, that is $O(N^2)$. Such a difference scheme is said to be *economical.*

The difference scheme (8), (9) is absolutely stable in some natural norms. If the solution $u(x, y, t)$ of the problem (1), (2) is sufficiently smooth on the closed parallelepiped $\bar{R}$, then the solution v of the difference scheme (8), (9) converges to u in the following sense:

$$\|u - v\|_h = O(h^2 + \tau^2), \quad (14)$$

where $\|u - v\|_h = \max\limits_{0 \leqslant \nu \leqslant M} \left(h^2 \sum\limits_{k,m=1}^{N-1} (u_{km}^{\nu} - v_{km}^{\nu})^2\right)^{1/2}$.

Sec. 34. The Dirichlet Problem for Poisson's Equation

Let $D = \{(x, y): 0 < x < 1, 0 < y < 1\}$ be an open square, Γ be its boundary, $\bar{D} = D \cup \Gamma$ be a closed square, and let $f(x, y)$ be a sufficiently smooth function defined on $\bar{D}$.

The Dirichlet problem consists in the following. It is required to find the function $u(x, y)$ continuous on $\bar{D}$ which on the open square D satisfies Poisson's equation

$$-\Delta u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(x, y), \tag{1}$$

and vanishes on the boundary of the square, that is,

$$u = 0 \text{ on } \Gamma. \tag{2}$$

The Dirichlet problem (1), (2) has the unique solution $u(x, y)$.

Let us set $h = 1/N$, $x_k = kh$, $y_m = mh$, $f_{km} = f(x_k, y_m)$ and construct the grids

$$\omega_h = \{(x_k, y_m): k, m = 0, 1, \ldots, N\},$$

$$\omega_h' = \{(x_k, y_m): k, m = 1, 2, \ldots, N - 1\},$$

$$\omega_h^* = \omega_h \setminus \omega_h' \ (\omega_h^* \text{ is the set of nodes lying on } \Gamma).$$

We then specify the norms

$$\|v\|_h = \max_{\omega_h} |v_{km}|, \quad \|v\|_h' = \max_{\omega_h'} |v_{km}|.$$

The Difference Scheme. It has the form

$$\Lambda v_{km} = f_{km}, \quad k, m = 1, 2, \ldots, N - 1, \tag{3}$$

$$v_{km} = 0 \quad \text{on} \quad \omega_h^*, \tag{4}$$

where Λ is the operator (33.5).

Written in a more detailed notation, the difference equation (3) is as follows:

$$-\frac{v_{k-1,m} - 2v_{km} + v_{k+1,m}}{h^2} - \frac{v_{k,m-1} - 2v_{km} + v_{k,m+1}}{h^2} = f_{km}. \tag{3*}$$

Its pattern is given in Fig. 26.

The solution v of the Dirichlet difference problem (3), (4) is found by the method of successive approximations using the scheme of variable directions (33.8), (33.9), where $f_{km}^{\nu - 1/2} = f(x_k, y_m)$, v_{km}^0 being arbitrary.

It is possible to prove that

$$\lim_{\nu \to \infty} v_{km}^{\nu} = v_{km}, \quad k, m = 1, 2, \ldots, N-1,$$

for any initial approximations v_{km}^{0}, the maximum rate of convergence being achieved for $\tau \approx h/\pi$. Here, we take as the basis the idea of the convergence, for $t \to +\infty$, of the solution of the heat equation to the solution of Poisson's equation if f is independent of t.

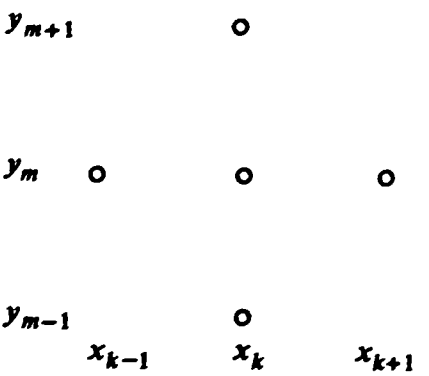


Fig. 28

The difference scheme (3), (4) is stable with respect to the right-hand side, that is, the difference problem

$$\Lambda z_{km} = \xi_{km}, \quad k, m = 1, 2, \ldots, N-1,$$
$$z_{km} = 0 \text{ on } \omega^*$$

has the only solution z for any $h = 1/N$, $N \geqslant 2$, this solution satisfying the inequality

$$\|z\|_h \leqslant c\|\xi\|_h', \tag{5}$$

where c is a constant independent of h and the grid function ξ.

Suppose that the solution of the Dirichlet problem (1), (2) is sufficiently smooth on the closed square $\bar{D}$, namely, $u(x, y) \in C_4(\bar{D})$. Then the difference equation (3) approximates the differential equation (2) on the solution u of the problem (1), (2) with second order with respect to h, that is,

$$\|\psi\|_h' = O(h^2), \tag{6}$$

where

$$\psi_{km} = \Lambda u_{km} - f_{km} \tag{7}$$

is the discrepancy of the solution u for the difference equation.

When obtaining the estimate (6), we take advantage of the fact that the second difference derivatives which on the basis of (10.3) approximate the indicated partial derivatives to within $O(h^2)$ correspond to the partial derivatives u''_{xx} and u''_{yy} entering into the equation (1). In Secs. 30 and 31 the analogous estimates of discrepancies are given in more detail.

Since, according to (4), the boundary condition (2) is precisely approximated on the grid ω_h^*, from (6) and the stability of the difference scheme (3), (4) with respect to its right-hand side there follows the convergence of its solution v to the solution $u \in C_4(\bar{D})$ of the problem (1), (2) with second order with respect to h, that is,

$$\|u - v\|_h = O(h^2). \tag{8}$$

Indeed, from the equation (3), equality (7), and conditions (2) and (4) there follows that the error $r = u - v$ on the grid ω_h is the solution of the difference problem

$$\Lambda r_{km} = \psi_{km}, \quad k, m = 1, 2, \ldots, N - 1,$$
$$r_{km} = 0 \text{ on } \omega_h^*.$$

Combining this with (5) and (6), we obtain (8).

The difference scheme (3), (4) has the second order of accuracy.

The Case of an Arbitrary Domain. Consider the Dirichlet problem

$$-\Delta u = f(x, y) \text{ on } G, \tag{9}$$

$$u = \varphi \text{ on } \Gamma, \tag{10}$$

where G is a finite domain (Fig. 27), Γ is the boundary of the domain G, $f(x, y)$ is the function defined on the domain G, and φ is the function specified on the boundary Γ.

As before, a square grid with step h is constructed. At all the nodes of the grid lying in the domain G which can be connected with four nearest nodes by means of line segments without intersecting the boundary Γ, the difference equation is given in the form

$$\Lambda v_{km} = f_{km}, \tag{11}$$

where Λ is the operator (33.5). The indicated nodes are shown in Fig. 27 by small circles. The pattern of the difference equation (11) is given in Fig. 26.

At the nodes lying in the domain G near its boundary Γ (marked in Fig. 27 with small triangles), difference equations are specified by the linear interpolation in the direction of the x- or y-axis. For instance, at the

point with the number 0 the equation has the form

$$v_0 = \frac{\rho_1}{\rho_1 + \rho_2} v_2 + \frac{\rho_2}{\rho_1 + \rho_2} \varphi_1, \tag{12}$$

where ρ_1 is the distance from the point 0 to the point 1 on the boundary Γ at which the given value of the function φ is taken, this value being denoted by φ_1; v_0 and v_2 are unknowns at the points 0 and 2, respectively; $\rho_2 = h$ is the distance between these points. For the sake of simplicity, here one index is used. The formula (12) means the linear interpolation between the points 1 and 2 into the point 0 (see Sec. 5). The difference

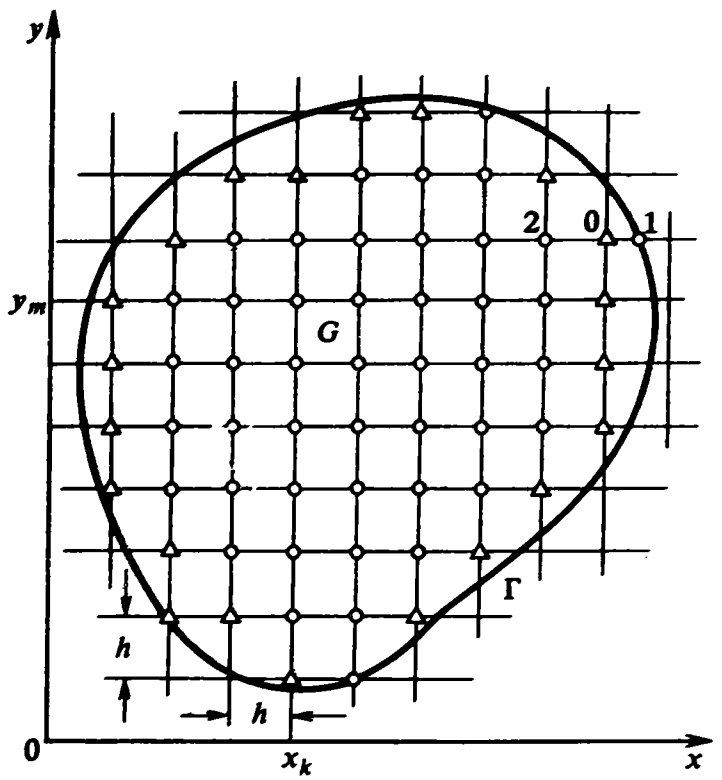


Fig. 27

equations at the remaining nodes marked with small triangles are specified in a similar manner. In this case, the distances from the point for which the interpolation is performed to both extreme points should not exceed h, and one or both extreme points must lie on the boundary Γ.

Solving the equations (11) given in a more detailed notation (3*), we have

$$v_{km} = \frac{v_{k-1,m} + v_{k+1,m} + v_{k,m-1} + v_{k,m+1}}{4} + \frac{h^2}{4} f_{km}. \quad (13)$$

Thus, at each node marked with a small circle the equation (13) is specified, and at each node indicated by a small triangle the equation has the form (12). The total number of equations coincides with the number of unknowns. The system of linear algebraic equations thus obtained has the unique solution v which can be found by applying the method of simple iterations and Seidel's method set forth in Sec. 21.

If the solution of the Dirichlet problem (9), (10) is $u(x, y) \in C_4(\bar{G})$, then the following estimate is valid:

$$\max_{G_h} |u - v| = O(h^2), \quad (14)$$

where G_h is the set of all the nodes marked with small circles and triangles. The solution $u(x, y)$ belongs to the class $C_4(\bar{G})$, for instance, if the boundary Γ has a three times continuously differentiable curvature, then the function φ of arc length s of the boundary Γ has a bounded fifth derivative, and $f(x, y) \in C_3(\bar{G})$.

Subject Index

Approximation of functions, 66
  best uniform, 67
  mean-square, by algebraic polynomials, 79
  mean-square, by trigonometric polynomials, 85
  uniform, 66
Axiom(s),
  of distance, 14
  of norm, 15
  of scalar product, 73

Best uniform approximation, 67
Buniakowski's inequality, 162

Choice of an optimal step, 58
Closed ball, 173
Coefficient(s),
  Fourier, 81
  Lagrangian, 32
Collocation points, 189
Comparing the Runge-Kutta methods with the Adams method, 129-130
Complete eigenvalue problem, 160
Convergence,
  in the mean, 17
  in the mean-square sense, 17
  in metric, 17
  in norm, 17
  uniform, 17
Counting the number of arithmetic operations performed, 139-141

Derivative(s),
  central difference, 55
  difference, 55
  second difference, 55
Difference(s),
  divided, 47
    of the first order, 47
    of order *n*, 47
    of zeroth order, 47
  finite, 45
    of the first order, 46
    of the *n*th order, 46
    of the second order, 46
Difference operator(s), 196
Difference scheme(s), 197
  absolutely stable, 224
  coditionally stable, 224
  for the differential boundary-value problem, 198
  economical, 229
  explicit, 221, 227
  implicit, 221, 227
  stable, 200
  unstable for any ratio of steps, 216
Digit(s),
  correct, 22
  significant, 22
Discrepancy, 189

Distance,
  axioms of, 14
  mean-square, 78

Element(s),
  first principal, 144
  leading, 134
  second principal, 144
Equation(s),
  heat, in two spatial variables, 227-229
  linear, with first-order partial derivatives, 209-216
  solving a system of, 173
  wave, 224-226
Error(s), 9
  absolute, 19
  of approximation by a spline, 65
  computational, 11
  of a function, 21
  inherent, 9
  interpolation, 33
  limiting absolute, 19
  limiting relative, 19
  of mean-square approximations, 88
  of the method, 10
  principal part of, 115
  relative, 19
  round-off, 11
Evaluating a polynomial, 26
Extrapolation, 30

Formula(s),
  Gaussian, 112
  parabolic, 104
  quadrature, 99-113
  rectangular, 101, 108
  Simpson's, 103, 108
  Simpson's composite, 107
  trapezoidal, 102, 108

General error estimate, 59
Gram determinant, 75
Gramian, 75
Grid function, 195

Horner's method, 26

Inequality,
  Buniakowski's, 162
  Minkowski's, 14
  triangle, 14
Interpolation, 30
  with equally spaced points, 42
  linear, 35
  phase, 35
Investigating stability by the perturbation method, 214

Jacobian matrix, 183

Kronecker-Capelli theory, 158

Leading element, 134
Lipschitz condition, 167

Method(s),
  Adams, 128
  collocation, 189
  of dividing a line segment into two equal parts, 184-185
  Euler's, 124
  of finding polynomials close to the best ones, 70
  Galerkin's 192
  Gauss elimination, 134
    backward procedure, 135
    forward procedure, 135
  grid, 226
  Horner's, 26
  improved Euler's, 127
  iteration, 168
  of least squares, 187
  of minimizing the discrepancy, 187
  Monte Carlo, 118-122
  Newton's, 180
  Newton's simplified, 183

numerical, of linear algebra, 133-166
predictor-corrector, 126
predictor-corrector, 126
Runge-Kutta, 126
Seidel's, 154
of simple iterations, 150
of specifying the inclinations of interpolation cubic splines, 63
of steepest (gradient) descent, 185
of straight lines, 226
of subdomains, 192
sweep, 156-160
of tangents, 180
Metric, 14

Newton's interpolation polynomial(s), 49-53
for backward interpolation, 52
for forward interpolation, 51
for unequal intervals, 50
Norm,
axioms of, 15
consistent, 146
Numerical differentiation, 53
formulas of, 55
simplest formulas of, 54

Optimal interpolation points, 39

Polynomial(s),
of best mean-square approximation, 77
of best uniform approximation, 67
criterion for, 68
Chebyshev, 37
properties of, 37
generalized, 77
interpolation, 30
Lagrange's interpolation, 32, 56
least deviating from zero, 38
Legendre, 82
Newton's interpolation, 49-53, 59
orthogonal, 81
Tailor's, 28
trigonometric, of order m, 85
Principal theorem of the theory of difference schemes, 202
Problem(s),
Cauchy's, 123
Dirichlet, for Poisson's equation, 230
partial eigenvalue, 160-166

Qadrature formulas, 99-113
canonical, 105
composite, 105
composite Gaussian, 113
defined by weights and nodes, 100
Gaussian, 109

Relation between the iteration method and Newton's method, 183
Richardson's refinement of an approximate solution, 117
Rounding off, 24
Rule(s),
for rounding off numbers, 24
Runge's, 116, 130, 207
for a practical error estimate, 114-118
for writing approximate numbers, 22

Scheme(s),
difference, economical, 229
with the main element chosen, 143
two-layer, 219
of unique division, 143
of variable directions, 228
Space,
Euclidean, 73
linear, 14
linear normed, 15
metric, 14
Spline(s), 62
cubic, 62
defect of, 62

defect of, 62
degree of, 62
inclination of, 62
Stability, 221
System of elements,
linearly dependent, 15
linearly independent, 15

Theorem(s),
Chebyshev's, 68
generalized mean-square, 99
Triangle inequality, 14

Uniform approximation of functions, 66
Unique division scheme, 143
Unstable difference scheme, 216

Variable directions, 228

Wave equation, 224-226
Writing approximate numbers, rule for, 22

Zero element, 14

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book.

We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics

Other MIR Titles

Ya. S. Bugrov, S. M. Nikolsky

HIGHER MATHEMATICS
Fundamentals of Linear Algebra and Analytical Geometry
Differential and Integral Calculus
Differential Equations. Multiple Integrals. Series.
Theory of Functions of a Complex Variable
A Collection of Problems

These three textbooks constitute a course in higher mathematics for engineering students and forming a four-book series entitled "Higher Mathematics". The collection of about 1200 problems has been compiled for the textbooks, thus completing the course.